RepliCueAuth: Validating the Use of a lab-based Virtual Reality Setup for Evaluating Authentication System

RepliCueAuth: Validating the Use of a lab-based Virtual Reality Setup for Evaluating Authentication Systems

Florian Mathis
florian.mathis@glasgow.ac.uk
University of Glasgow
Glasgow, United Kingdom

Kami Vaniea
kvaniea@inf.ed.ac.uk
University of Edinburgh
Edinburgh, United Kingdom

Mohamed Khamis
mohamed.khamis@glasgow.ac.uk
University of Glasgow
Glasgow, United Kingdom

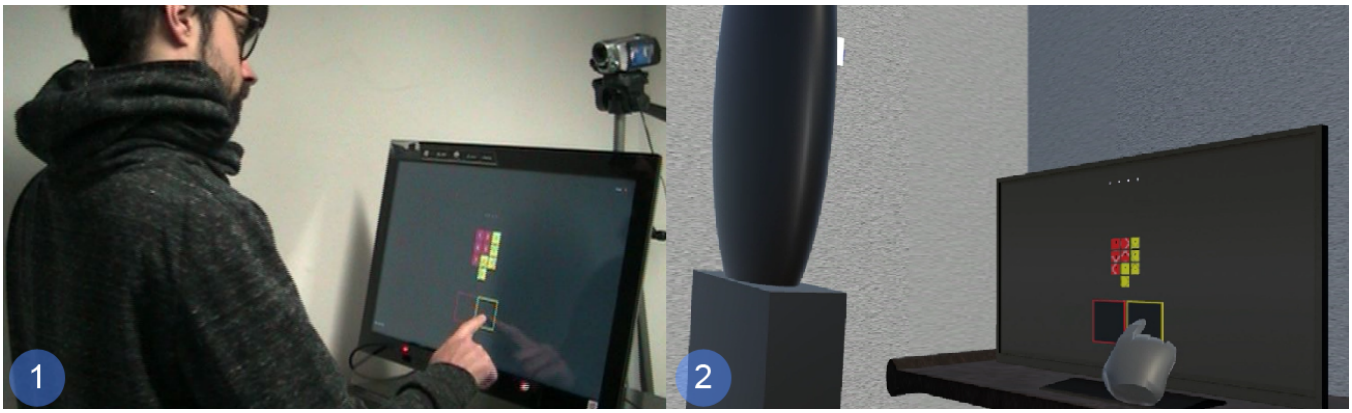

Figure 1: We evaluate and discuss the suitability of using Virtual Reality (VR) to conduct human-centred usability and security evaluations of real-world authentication systems. To this end, we replicated a recently introduced authentication scheme called CueAuth [52] (❶) into VR (❷). We then evaluate the usability and security of our replica and compare the results to the real-world evaluation of CueAuth [52].

ABSTRACT

Evaluating novel authentication systems is often costly and time-consuming. In this work, we assess the suitability of using Virtual Reality (VR) to evaluate the usability and security of real-world authentication systems. To this end, we conducted a replication study and built a virtual replica of CueAuth [52], a recently introduced authentication scheme, and report on results from: (1) a lab-based in-VR usability study (N=20) evaluating user performance; (2) an online security study (N=22) evaluating system's observation resistance through virtual avatars; and (3) a comparison between our results and those previously reported in the real-world evaluation. Our analysis indicates that VR can serve as a suitable test-bed for human-centred evaluations of real-world authentication schemes, but the used VR technology can have an impact on the evaluation. Our work is a first step towards augmenting the design and evaluation spectrum of authentication systems and offers ground work for more research to follow.

CHI '21, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05…$15.00
https://doi.org/10.1145/3411764.3445478

CCS CONCEPTS

• **Human-centered computing** → **Virtual reality**; • **Security and privacy** → **Usability in security and privacy**.

KEYWORDS

Virtual Reality, Research Method, Usable Security, Authentication

ACM Reference Format:
Florian Mathis, Kami Vaniea, and Mohamed Khamis. 2021. RepliCueAuth: Validating the Use of a lab-based Virtual Reality Setup for Evaluating Authentication Systems. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3411764.3445478

1 INTRODUCTION

Usable Privacy and Security (USEC) researchers have proposed a plethora of novel authentication schemes (e.g., [18, 19, 40, 52, 60, 105]). Developing and evaluating prototype authentication systems often involves non-commodity hardware (e.g., special smartphone prototypes [18], private near-eye displays [111], or eye trackers [49]), and complex study setups (e.g., [7, Figure 3], [50, Figure 2]). This makes corresponding usability and security evaluations often costly and time consuming. While there are infrastructures that allow running online studies (e.g., Amazon Mechanical Turk) and result in valuable and inspiring privacy and security research [7, 70, 84, 85], they are often not suitable for USEC research involving physical prototype systems. A promising emerging evaluation paradigm is Virtual Reality (VR) studies [64, 89, 104]. VR studies

allow researchers to conduct evaluations by having participants experience virtual replicas of real-world systems in an immersive VR environment. VR studies can (a) reduce costs of studies evaluating real-world authentication systems as researchers do not need to build physical prototypes; (b) allow recruiting many and diverse participants remotely, thereby increasing ecological validity in the long run; and (c) reduce the need for face-to-face studies which could be advantageous at times e.g., during pandemics.

However, we are a long way from being able to apply VR studies for usable privacy and security research. As a first step, we aim to investigate whether and how VR studies can complement usability and security evaluations of real-world authentication systems. For example, are we able to learn about users' performance on a real-world system by measuring their performance on a virtual reality replica? Or does an authentication scheme's vulnerability to observation attacks when used in VR map to similar weaknesses if used in the real world?

Determining which results from lab-based VR usability and security studies match those obtained from the real world can be particularly valuable. If USEC researchers could quickly iterate and evaluate their authentication prototypes in VR before real-world deployment, they could save costs, time, effort, and could recruit potentially larger and more diverse samples in the long run.

To see if a real-world study of an authentication scheme can be similarly run in VR, we set out to do an alternative methods replication study of CueAuth [52]. To accomplish this, we replicated CueAuth in VR and evaluated it using two in-depth studies: 1) an in-lab VR *usability study* (N=20) using VR replicas of the authentication approaches used in the CueAuth paper as well as similar metrics and 2) an online *security study* (N=22) to study the observation resistance of our VR replica under two threat models as done in CueAuth [52]. We then compared the results from our studies with the real-world results of CueAuth [52]. Our studies and the comparisons with earlier work allow us to draw novel insights about the strengths and weaknesses of VR studies for usability and security evaluations of authentication schemes. The results suggest that many findings are transferable to the real world. Measures like entry accuracy, and users' experience and perception of input methods are largely similar across our lab-based VR studies and the original real-world studies [52]. Perceived workload when providing input with touch and mid-air remained similar across both study types, but not for eye gaze. Input with eye gaze was perceived as less demanding in VR. A difference was also found in users' entry time when using touch, indicating that artificial artefacts such as virtual hands can negatively impact users' performance when the technology is not mature enough to provide experiences similar to reality. The other difference was found for eye gaze where input was significantly faster in VR than in the real-world study because of the better eye tracking conditions. Inline with the real-world security study of CueAuth, observations against VR avatars during authentications are more successful in threat models that involve video recordings. Observation attacks on touch input are also more accurate than on mid-air and eye gaze. We conclude by discussing our validation and how VR studies could be applied to more general usable privacy and security research in the future.

1.1 Contribution Statement

The contribution of this paper is threefold: (1) We propose the idea of using Virtual Reality as a test-bed for usability and security evaluations of real-world authentication systems. (2) We complement prior work that evaluated usability aspects in VR [64, 104] by the first lab-based in-VR usability evaluation and the first online security evaluation through recordings in VR of a real-world authentication system and validate the use of VR through a comparison with the real-world study [52]. (3) Finally, we derive lessons learned to support researchers in designing, developing, and evaluating authentication systems of similar type in VR and discuss potential follow-up research directions.

2 RELATED WORK

To contextualise our work, we review evaluation methods and revisit prior work focusing on VR as a study platform.

2.1 Empirical Evaluations

When it comes to the evaluation of research artefacts, researchers apply different methods. There is an increasing consensus of accepted evaluation methods within the CHI community [9], with lab studies being still the most popular evaluation method [11, 58]. Empirical evaluations span from small-scale lab studies [1, 20, 87] to large-scale in-the-wild evaluations [15, 25, 106]. Lab studies are suitable for evaluations in controlled settings to iteratively evaluate systems in a cheap way [23]. A drawback is that they often do not represent a natural context, resulting in a low external validity [23, 37]. There are ongoing debates whether or not conducting in-the-wild studies is worth the hassle to increase the external validity [55, 56]. Although they are often considered to be the way-to-go when aiming for evaluations in a natural environment [23], they are often expensive and time-consuming [56]. Moreover, there are many discussions around legal and ethical burdens when conducting this type of research, especially within privacy and security research [26, 69]. Many USEC researchers therefore draw on online studies to get to scale and increase external validity [25, 85]. There is also a notion of an increase of using online platforms such as Stack Overflow or PatchManagement.org to conduct security and privacy research [44, 98]. The variety of these evaluation methods shows, as Greenberg & Buxton put it, that there is no all-in-one solution and evaluations are not "universal panaceas" [112].

2.2 Virtual Reality as an Evaluation Method in Human-computer Interaction

HCI researchers recently started to look into alternative evaluation methods to cope with different study requirements and challenges. Recent work explored behaviour in front of public displays in VR, and compared it to real-world behaviour to find many similarities [64]. There is also work that compared conducting empirical studies online, in VR, in AR, in the lab, and in in-situ studies to find that some findings are comparable across the methodologies while responses to standardised questionnaires such as AttrakDif [39] and ARI [30] yielded significantly different results [104]. Others compared navigation methods in VR to the real world to find differences in e.g., navigation performance while there was no difference in users' route recognition rate across the two study types [89], or

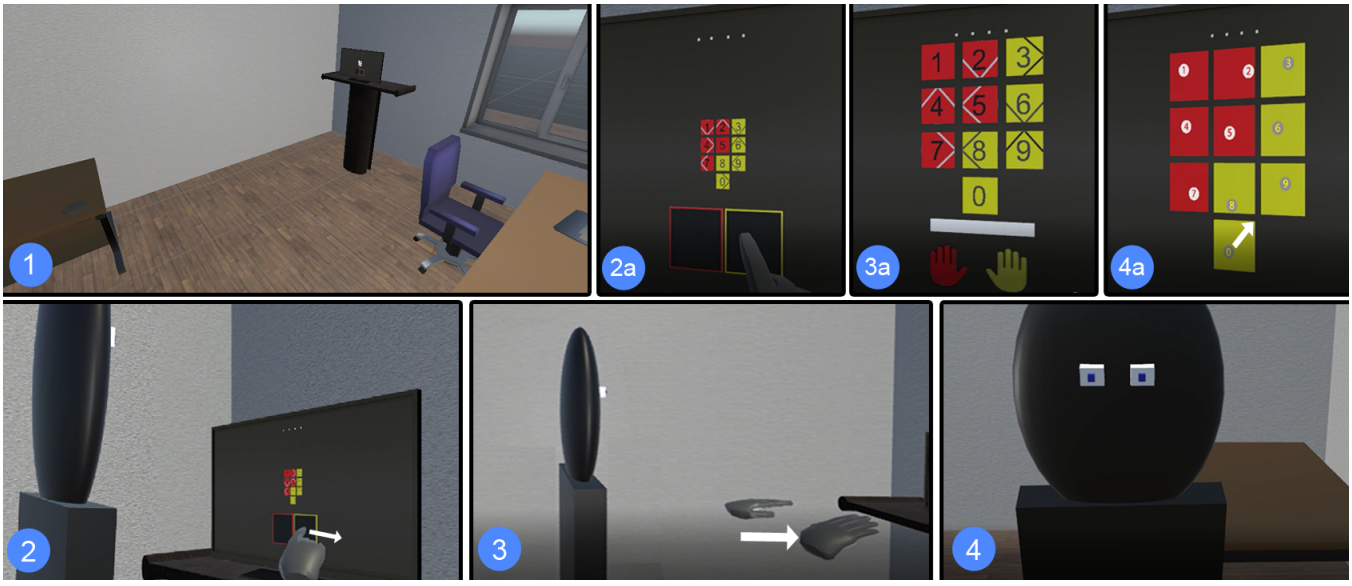

Figure 2: We replicated CueAuth's setup into a VR room with a situated display that shows PIN-pads (❷-a,❸-a,❹-a) featuring the cues [52]. To select a digit, the user responds to the cue displayed on its button. After selection, the cues are randomly reshuffled. To enter "0" via touch in the shown example, the user makes a touch gesture to the right in the yellow box to correspond to the yellow "0" button with a right-arrow (❷). To enter "1' in mid-air, the user performs the gesture to the front with their left (red) hand as the "1" button is red and has no arrows (❸). In eye gaze, each digit moves along a distinct trajectory, allowing users to follow the movement with their eyes to make selections. To enter "0", the user has to follow the diagonal movement of the digit with their eyes (❹).

studied teaching and learning classical orientation and mobility tasks for visually impaired people in VR [101]. There are several additional works (e.g., [33, 73, 75, 91]) that emphasise the potential of using VR as a research platform.

2.3 Lessons Learnt from Prior Work

Prior work asserts that there is no "Swiss Knife" study method. The choice of study method should in general evolve from the actual problem and the research questions [34, 92]. VR has the potential to act as a research paradigm to complement widely-used study methods such as lab, online, and field studies. Prior work also suggests that users behave similarly in VR as they do in the real world [64, 73, 104], indicating a high transferability of findings collected in VR to the real world. Yet, to date, it is unclear to what extent results of usability and security evaluations of authentication schemes fully conducted in VR are transferable to the real world. While previous work focused mostly on comparing user behaviour in front of public displays in VR and the real world [64] or assessed smart artefacts through standardised questionnaires [104], less is known about the transferability of quantitative measures such as entry accuracy, entry time, and observation resistance from VR to the real world. A validation of the use of VR for evaluating authentication schemes allows the USEC community to leverage this opportunity and identify when it is useful to employ VR studies.

3 RESEARCH PREFACE

Research in USEC covers a wide range of areas, many of which are not necessarily easy to create in-lab experiments for. Our work focuses on authentication because it is a major theme in USEC research [27] and novel authentication scheme research often involves physical prototypes (e.g., [18, 71]) as well as complex study setups to asses a system's security (e.g., [7, Fig. 3], [47, Fig. 2]). We replicated CueAuth [52] because: (1) it covers a range of input methods that are used in security research (touch [19, 107], mid-air [1, 6], and eye gaze [20, 46, 61]); (2) it provides a holistic usability and security evaluation; and (3) its underlying concept has already been studied in different contexts [52, 105, 108].

3.1 CueAuth: An Overview

CueAuth [52] is an authentication system on situated displays such as vending machines, ATMs, and other public displays in social spaces. Providing input on CueAuth [52] is considered to be fast and highly secure against observations. To enter a PIN, users either perform touch gestures, mid-air gestures, or smooth pursuit eye movements [103]. The underlying concept of CueAuth [52] is based on cues on the screen (Figure 2-❷a,❸a,❹a). For touch and mid-air, arrows on the respective digits show the users which gestures they have to provide to enter the corresponding digit. The absence of an arrow indicates that users have to tap (in touch) or perform a gesture towards the front (in mid-air). In eye gaze, CueAuth [52] employs Pursuits [103], a calibration-free gaze interaction method. Smooth pursuit eye movements are compared to the trajectories of animated targets (i.e., digits 0 - 9) to determine which digit users gaze at. In all three input methods, the cues are randomly reshuffled after every input and all input concepts are inline with the ones

used in the original real-world evaluation of CueAuth [52]. We describe the concept of each input method below.

3.1.1 Touch. To provide input via touch users need to observe which cue is shown on the digit (e.g., see Figure 2-2a) and then perform the corresponding touch gesture in the respective box, i.e., digits on the left are entered in the red box (i.e., 1,2,4,5,7) and digits on the right (i.e., 3,6,8,9,0) are entered in the yellow box. Note that an arrow means a touch gesture to the displayed direction (e.g., an arrow to the left means a swipe gesture to the left). To select the digit *"3"* in Figure 2-2a users need to make a down touch gesture in the yellow box.

3.1.2 Mid-Air. In mid-air, users raise their hands and select digits via mid-air gestures in the direction of the corresponding arrow (e.g., see Figure 2-3a). Similar to touch, the gestures are performed with the left hand if the digit is coloured red and with the right hand if the digit is coloured yellow. To enter the digit *"3"* in Figure 2-3a users need to make a gesture to the right with their right hand.

3.1.3 Eye Gaze. Contrary to touch and mid-air where users perform gestures based on the arrows, in eye gaze users need to follow moving targets with their eyes (Pursuits [103]). The advantages of Pursuits [103] over location-based gaze gestures are manifold. For example, the input does not require accurate gaze estimation; thus, does not require eye tracker calibration, which reduces the pre-interaction time to a minimum. Pfeuffer et al. [79] showed that calibration impacts the usability and user experience in a negative way when interacting with public displays as such interactions are often rather of shorter duration [74]. To, for example, select the digit *"0"* in Figure 2-4a users need to follow the diagonally moving target with their eyes.

3.2 Overview of Studies

Similar to the original real-world study of CueAuth [52], we used a repeated measures design for both the usability and security study. Conditions were counter-balanced using a Latin Square. We recruited new participants for each study — no participant took part in more than one study. Both studies are designed as conceptual replications using "alternative methods" [110]. The in-VR usability study was conducted in-person using equipment we provided to ensure consistency of study environment and protocol between participants. In future work we would be interested to see how consistent the results are with participant-owned equipment in their own homes as the increased adoption of new technologies means that many households will likely have access to VR in the near future [22], but we were concerned that doing so here might add unnecessary variance. We used in-VR questionnaires to ensure a consistent VR experience and not break participants' focus [5, 81], doing so also makes the methodology more applicable to potential future studies running fully remotely. The recruitment of the usability study considered the same user profiles as the original real-world study [52], i.e., users with normal/corrected-to-normal vision and no prior experience with cue-based authentication. We conducted the security study online through Prolific [80] using pre-recorded videos recorded in the VR environment. Prolific is an established platform for online subject recruitment for scientific purposes and considered to be a valuable alternative to other crowdsourcing platforms and is regularly used for advertising academic studies in HCI and USEC (e.g., [3, 66]). We embedded recordings of authentications in VR (e.g., see Figure 2-3/3a) in Qualtrics, an online survey tool that can be accessed via web browsers. This depicts the same procedure used in the original real-world study [52] where participants were showed video recordings of authentications in the real world, but in our case we used video recordings of authentications performed in the virtual environment through a virtual avatar. Using recordings of authentications to assess a system's resistance to observations is a commonly used approach in usable privacy and security research [8, 18, 19, 47]. We applied a pre-screening to ensure that participants of the online security study have English proficiency and access to a PC. Finally, the results from the usability and security study were compared with the real-word results observed in the CueAuth paper [52].

3.2.1 Ethics and Compensation. Our studies were approved by the College of Science and Engineering Ethics Committee at the University of Glasgow and consent was obtained prior to both studies. We compensated participants €15.00 for the usability study and £7.50 for the security study. Participants of the security study also took part in a draw to receive an additional £7.50 based on their observation performance. This compensation method was also used in CueAuth [52] and is commonly used to motivate participants in security studies [29, 67, 68]. Participants could optionally share photos of any notes they took during the security study for an additional compensation of £0.5.

3.3 Apparatus and Implementation

We implemented our VR prototype using Unity3D C#, Leap Motion SDK [72] for the finger tracking, and Tobii XR SDK [102] for eye tracking. We used the HTC VIVE Tobii DEV KIT [100] which we connected to a VR-ready laptop *(Razer Blade 15, NVIDIA GeForce RTX 2080)* [83]. The used VR headset comes with an integrated Tobii eye tracker (120 Hz).

3.3.1 Implementation. We aimed for similar implementations as used in the original study [52]. Due to the nature of VR some implementations differ, but the overall concepts remain the same.

Touch. Instead of calculating the distance between on-screen touch points to detect touch gestures as done in CueAuth [52], we used colliders and the OnCollisionEnter, OnCollisionStay, and OnCollisionExit event listeners [99] around user's touch point. One collider was positioned at user's initial touch point and the others (left/right/top/bottom) ≈3.5 cm away; this value is based on pilot tests. A touch gesture is registered depending on which collider the user's finger collides with. If none of the colliders on the side are touched, but the touch exits the collider at user's initial touch point, the system recognises a tap. Instead of providing haptic feedback to simulate touching a screen in VR through, for example, electrical muscle stimulation [63] or haptic gloves [43], we location-mapped a physical surface to the in-VR screen to make users touch the physical surface when performing touch gestures on the virtual surface of the screen (Figure 3). This is inspired by the work of Kim et al. [54].

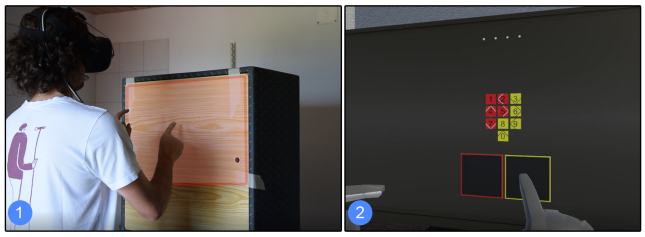

Figure 3: We mapped a physical surface in the real world (❶) to the virtual screen (❷) in the virtual environment. Following this approach provides users with haptic feedback when touching the screen without using an actual touchscreen.

Mid-Air: Instead of tracking mid-air gestures through an external device as done in the real-world study with a Microsoft Kinect One, we attached two HTC VIVE trackers to users' wrist. The default position is where users' hands are raised and parallel to the elbows (see virtual avatar in Figure 2-3). A small threshold area (10 cm, determined through pilot tests) around the default position was defined as "no-input area". Gestures were detected using colliders the same way we did for touch. After each gesture, users' hands had to return to the default position before the next input.

Eye Gaze: As done in the original CueAuth paper [52], we use the implementation by Vidal et al. [103] to detect smooth pursuits. A moving digit, used as stimulus for Pursuits [103], is selected if the correlation between its trajectory and the user's eye movements exceeds a Pearson correlation coefficient threshold. The stimulus with the highest correlation above the pre-defined threshold to the user's eye movements is defined as the stimulus at which a user gazes at. The threshold (>0.8) as well as the trajectories of the stimuli (circular, linear diagonal, and zigzag) are based on the original study [52]. Different configurations of thresholds could lead to different results in terms of entry accuracy and entry speed [103] – we therefore did not fine-tune the threshold and used the same as in the original real-world study [52].

3.4 Statistical Analysis and Data Visualisation

Our statistical analysis entails (1) an analysis of the repeated measures usability and security VR studies; and (2) a comparison between our VR experiment's results and those obtained from the original real-world study of CueAuth. The latter feeds our discussion of the validity of VR studies in USEC research. To visualise our data we use bar charts and violin plots [41], displaying a rotated kernel density plot on either side of a box plot that shows the mean and standard deviation. We colour-code the main observations related to comparing VR and real-world studies by green, orange, and red to denote a high (🟩), intermediate (🟧), and low (🟥) match in the results.

3.4.1 VR Study Analysis. In all our statistical analyses, we applied the same tests as used in the original work [52]. We use repeated measures ANOVAs (IV: input method) in the usability study and two-way repeated measures ANOVAs (IV: input method, threat model) in the security study. We checked our data for normality prior to the analysis. We applied ANOVAs to parametric and normal/near-normal distributed data as ANOVAs are robust to deviations from normality [32]. Post-hoc pairwise comparisons were Bonferroni corrected for controlling familywise errors. Greenhouse–Geisser adjustment was used to correct for violations of sphericity. For qualitative analysis we used a code book [88] based on the findings of CueAuth [52] to support comparison. We also added a VR code to capture VR-related comments that were obviously not present in the original real-world usability study. We kept an eye out for potential new codes, but did not observe any. The content was fairly simplistic, so one author did all the coding.

3.4.2 Validation Analysis. To assess the validity of our VR studies, we used between-group analysis to compare our studies to the real-world evaluation of CueAuth, whose full dataset was obtained through the original paper's first author [52]. We applied two-way mixed ANOVAs with one between-subjects factor (study type: VR vs real world) and one within-subjects factor (input method) in the usability study and three-way mixed ANOVAs with one between-subjects factor (study type) and two within-subjects factors (input method and threat model) in the security study. While this approach allows us to reveal significant differences between the two study types, a non-significant outcome (i.e., p-value > 0.05) does not indicate that the values are equal or there is no effect of study type on the measures [45]. The sample size in our VR studies is determined by the original CueAuth study [52]. Such sample sizes increase the likelihood of type-2 errors for statistical tests. Therefore, in addition to reporting non-statistically significant pairs between the real-world study and our VR study, we also report similar patterns across the two study types.

4 STUDY 1: IN-VR USABILITY STUDY

Our study design follows the real-world study of CueAuth [52] as much as possible. We had the input method as our only independent variable, with three levels: (1) touch; (2) mid-air; and (3) eye gaze.

4.1 Procedure and Task

Each participant went through three blocks in total, one per condition. Participants first filled in demographics followed by one of the three interfaces (Figure 2). Prior to each block, we explained the input method. Participants then performed training runs to become acquainted with the corresponding input method. We excluded all training runs from the analysis. Our system verbally announced each 4-digit PIN which we pre-defined in advance. Each participant had one chance to enter each PIN. After entering 16 PINs per condition, participants filled in in-VR questionnaires [4, 81]. The same was repeated for the other methods. We concluded with semi-structured interviews guided by the results from the original study [52].

4.2 Results

We recruited 20 participants (8 female, 12 male, self-reported) through social media, word-of-mouth, and local societies. Our participants were on average 27.25 years (range: 18 - 57, SD=8.31) and their demographics (gender, age) correspond roughly to the demographics of the participants in the real-world study (13 female participants, ages ranging from 18 to 33 years (M=24.1, SD=3.9) [52]). We measured entry accuracy, entry time, and the perceived workload using NASA-TLX [38]. We excluded the data of five participants (vs three

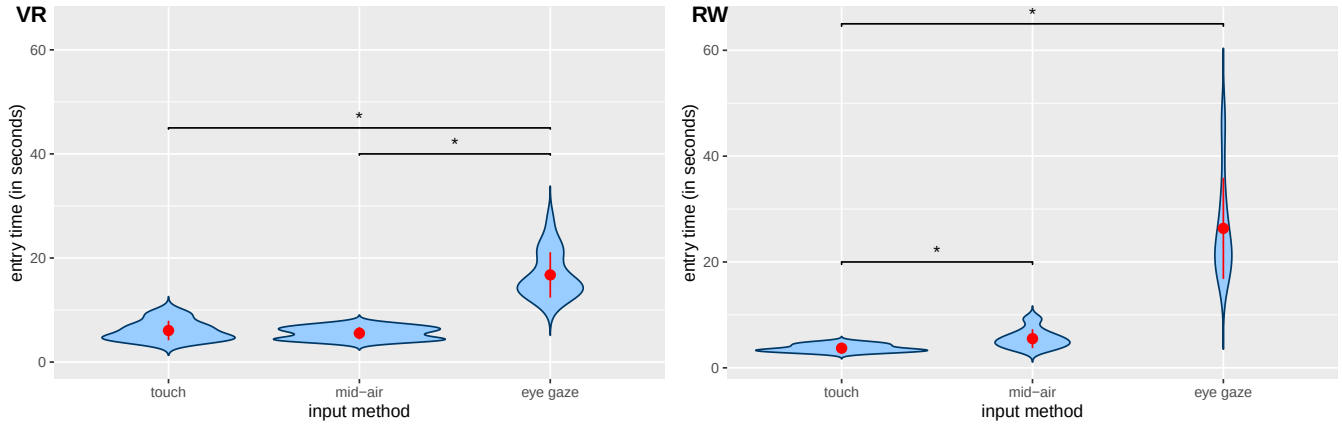

Figure 4: Users authenticate significantly faster when using touch or mid-air in the VR study compared to eye gaze. However, there is no evidence that touch in VR is faster than mid-air, or vice versa. This is slightly different from the real-world's study where touch was significantly faster than mid-air and eye gaze. Note that the red pointrange denotes mean ± standard deviation.

in the real-world study [52]) due to tracking issues. Three (P3, P4, P16) in the touch, and two (P14, P18) in the mid-air condition.

4.2.1 Entry Accuracy. We found a significant main effect of input method on entry accuracy, $F_{1.451,20.311} = 5.791$, $p<0.05$. Post-hoc pairwise comparisons revealed significant differences ($p<0.05$) in entry accuracy between touch (M=89.97%, SD=8.10%) and mid-air (M=80.42%, SD=6.19%). No significant differences were found between the other pairs. Entry accuracies are high for all methods with 89.97% (SD=8.10%) for touch, 83.75% (SD=9.68%) for eye gaze, and 80.42% (SD=6.19%) for mid-air. When comparing our results to the original real-world study, no statistically significant interaction (input method × and study type) was found in case of entry accuracy, $F_{2,62} = 0.401$, $p = 0.671$. There was also no main effect of study type on entry accuracy, $F_{1,31} = 0.058$, $p = 0.812$. However, while touch input was significantly more accurate than eye gaze in the real-world study [52], this was not the case in our VR study.

Observation #1 [Entry Accuracy]: We did not find any significant differences of entry accuracies between our VR study and the original real-world study. This means that there is no evidence that entry accuracy when providing input with touch, mid-air, and eye gaze differs between the two study types.

4.2.2 Entry Time. We found a significant main effect of input method on entry time, $F_{1.229,15.972} = 69.778$, $p<0.05$. Significant differences ($p<0.05$) were found between eye gaze (M=16.75 s, SD=4.36 s) and touch (M=6.06 s, SD=1.87 s), and between eye gaze (M=16.75 s, SD=4.36 s) and mid-air (M=5.54 s, SD=1.16 s). This means that PIN entries in touch are significantly faster than in eye gaze, which matches the results from the real-world study [52]. Whereas touch was also significantly faster than mid-air in the real-world study, this was not the case in our VR study. No significantly different entry times were found for mid-air and touch. Figure 4 shows the distributions. When comparing the entry times to the original real-world study, we found a significant interaction effect (study type × input method), $F_{1.054,31.614} = 13.908$, $p<0.05$. Follow-up analysis revealed a statistically significant difference in entry time when using touch between our VR and the real-world study, $F_{1,33} = 24.617$, $p<0.05$. Touch input was significantly faster in the original study (M=3.73 s, SD=0.98 s) than in our VR study (M=6.06 s, SD=1.87 s). The other was found for eye gaze, $F_{1,35} = 15.728$, $p<0.05$. Input with eye gaze was significantly faster in our VR study (M=16.75 s, SD=4.36 s vs M=26.35 s, SD=22.09 s). No significant difference between the study types was found for mid-air (M=5.54 s, SD=1.16 s vs M=5.51 s, SD=3.87 s).

Observation #2 [Entry Time]: Touch input was significantly faster in the original real-world study than in the VR study, whereas eye gaze was significantly faster in the VR study than in the real-world study. Entry time using mid-air remained the same across both study types.

4.2.3 Perceived Workload. We did not find any statistically significant differences of the mean raw NASA-TLX values between the input methods, $F_{2,38} = 0.389$, $p=0.681$. The overall task load indexes are 34.83 (SD=23.61), 34.0 (SD=19.68), and 30.33 (SD=18.33) for touch, mid-air, and eye gaze in VR. We ran multiple repeated measures ANOVAs on the level of each NASA-TLX dimension to investigate if there is an effect between the input methods on the level of each NASA-TLX dimension. A significant main effect was found for input method on performance, $F_{2,38} = 7.615$, $p < 0.05$. Post-hoc pairwise comparisons revealed a significant difference between eye gaze and touch ($p<0.05$) and eye gaze and mid-air ($p<0.05$). Figure 5-VR shows the mean scores and the pairs that are significantly different.

When comparing users' perceived workload in our VR study to the real-world study, a two-way mixed ANOVA revealed a significant interaction effect (study type × input method), $F_{2,76} = 7.233$, $p < 0.05$. There was a statistically significant difference in users' perceived workload when using eye gaze between the two study types, $F_{1,38} = 7.803$, $p < 0.05$. Follow-up analysis revealed a significant difference ($p<0.05$) between eye gaze in the two study types in terms

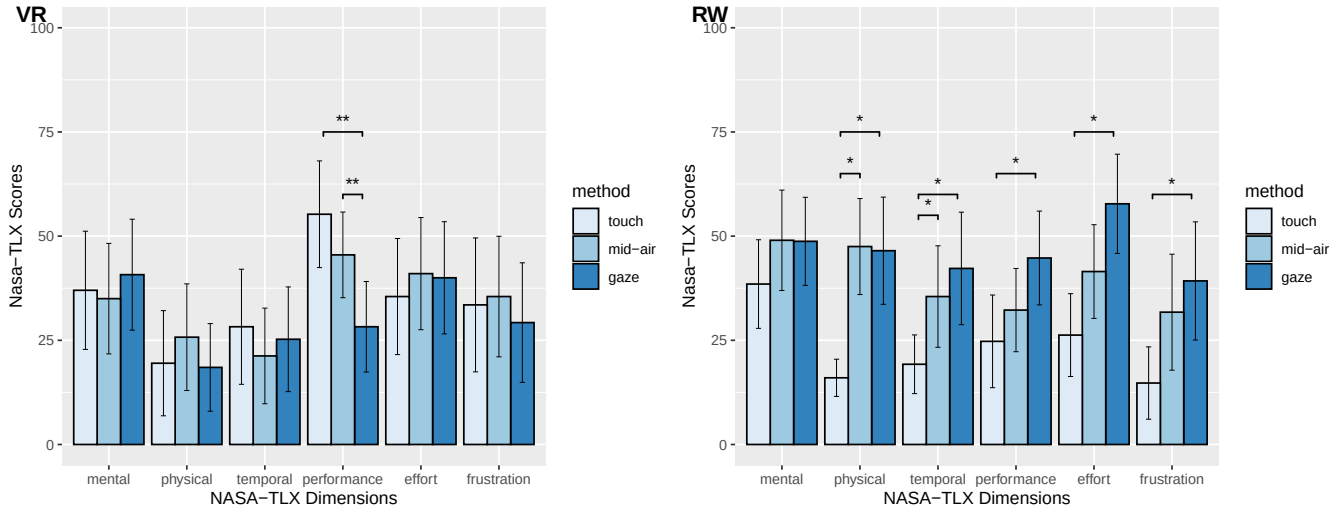

Figure 5: We did not find any significant differences within our VR study between the mean raw NASA-TLX values on the level of the input methods, indicating that users perceived all three methods equally demanding. However, analysing each dimension revealed a significant difference ($p<0.05$) between touch and eye gaze, and mid-air and eye gaze in terms of performance. Eye gaze in the real-world study (RW) was perceived as more demanding than in VR. Note that black lines denote the 95% confidence interval (CI).

of physical workload, frustration, and effort. For our VR study the mean raw values for physical workload, frustration, and effort are 18.50 (SD=21.83), 29.25 (SD=29.79), and 40.0 (SD=27.96). For the real-world study the values are 46.5 (SD=26.71), 39.25 (SD=29.44), and 57.75 (SD=24.73). We did not find any other significant differences.

Observation #3 [Perceived Workload]: There is no evidence that users' perceived workload differs in terms of touch and mid-air in both the real-world study and the VR study. However, participants self-reported significantly lower physical workload, less frustration, and less effort when providing input with eye gaze in VR.

4.2.4 Qualitative Feedback. We collected feedback through loosely guided semi-structured interviews (see Appendix A.1). We transcribed the interview data and used the code book [88] based on the findings in CueAuth [52] to tie the comments voiced in our VR study to the original real-world study. Additionally, we report on the perceived impact of VR on users' behaviour. Although a strict comparison of qualitative data is challenging [31], we can see many similarities.

Exposure to the Input Methods: Similar to comments voiced in the real-world study, the VR study participants also reported being exposed previously to touch input (e.g., smartphones) and mid-air gestures (e.g., Xbox video game console), but were less exposed to gaze-based interaction, which was experienced by only two participants before.

Perception of the Input Methods: Participants perceptions of the methods in VR matched those from the real-world study. While their ranking (see Section 4.2.5) suggests that they preferred eye gaze over mid-air and touch, they associated touch with more positive attributes than eye gaze. Examples include intuitive, realistic, and effortless. Although mid-air gestures were similarly positive perceived, there were also more negative attributes associated compared to touch. Mid-air gestures require more explicit movements than touch and look weird. P13 thought (*"[mid-air is] neither fish nor fowl"*). Gaze was perceived as long-winded and exhausting, but perceived safer than touch and mid-air. Input with mid-air and eye gaze was also described as hygienic.

Usability: We received mixed comments about the usability of each method. Touch was found simple and familiar. However, participants also voiced that providing touch gestures with virtual hands feels strange. While some participants perceived mid-air gestures as comfortable, others mentioned that it feels weird in public: *"looks like a jumping jack", P17.* Gaze was perceived as long-winded, but secure. This suggests that touch input was perceived slightly more usable than mid-air and eye gaze. However the additional technological layer (=the virtual hand) was taken negatively. This deemphasises the advantages of touch input that were dominant in the real-world study [52] due to VR components (Figure 6-VR/RW).

Enhancement: Participants voiced the lack of proper feedback when entering a PIN and suggested to extend the interaction space of mid-air gestures through finger or face gestures. Similar enhancements were also mentioned by participants in the original real-world study [52]. Others mentioned technological limitations in finger tracking and gesture detection. In terms of eye gaze, participants criticised the pixelated targets they had to follow with their eyes.

Perception of VR: Most participants voiced that the virtual environment did not impact the way they interacted with the system. However, others mentioned that they felt isolated in VR which allowed them to pay more attention to the task than they would

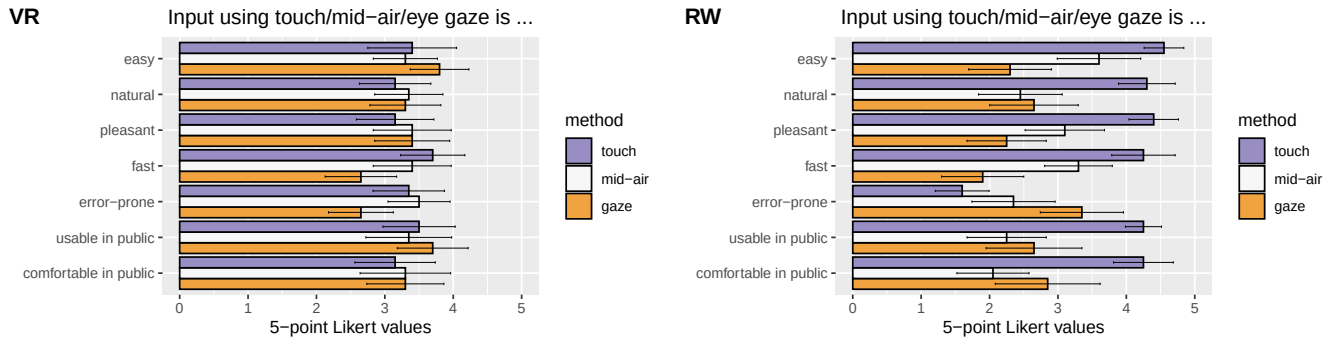

Figure 6: Participants in our VR study rated the three input methods on 5-point Likert scales, from 1 *(Strongly Disagree)* to 5 *(Strongly Agree). RW* shows the results from the original real-world study [52]. Note that black lines denote the 95% confidence interval (CI).

in the real world. P13 even said that she was totally unaware of experimenter's presence during the study. P19 mentioned that he was inhibited to make specific movements because of being afraid of bumping into real-world obstacles. P20's opinion was different. He voiced that he treated the VR environment as a safe space and therefore felt "freer" to perform gestures. Participants described people in the real-world as "additional noise" that is not present in VR. P6 said the room was *"too clean to be realistic"*.

Overall, our interviews represent a similar picture as participants' answers to our Likert questions (Figure 6). Compared to participants' qualitative feedback in the real-world study, touch input was perceived as, for example, more challenging and more error-prone in our VR study than in the real-world study. These results can be attributed to the finger tracking used in our VR study that we discuss further in Section 6.1.

Observation #4 [Qualitative Feedback]: Most feedback is similar across the VR and real-world study. However, the different sensing capabilities impacted participants' perception and preference of some methods.

4.2.5 Ranking. We asked participants to rank their preference of the three input methods. Raw scores were multiplied by their weight factor: ×3 for rank 1, ×2 for rank 2, ×1 for rank 3, and then summed up to compute weighted scores (based on [96]). Eye gaze was the most preferred one (45), followed by mid-air (41), and touch (37).

Observation #5 [Ranking]: In contrast to the real-world study where touch was the most preferred and eye gaze the least preferred input method, users in the VR study preferred eye gaze over mid-air and touch.

5 STUDY 2: REMOTE EVALUATION OF THE SYSTEM'S RESISTANCE TO OBSERVATIONS THROUGH VR RECORDINGS

For the security evaluation, we designed a study in which participants took the role of an attacker with the aim to attack 4-digit PINs entered by a virtual avatar (Figure 2). We used Qualtrics [82] and Prolific [80] to deploy the study online (see also Section 3.2).

5.1 Threat Models

We considered the two threat models used in the original real-world study [52]. In both models, the attacker knows how the system works and has an optimal view on user's input. **Single attack:** The attacker has only one chance to observe the authentication; **Repeated-video attack:** The attacker can watch a video recording of the authentication as often as they wish. They can pause, rewind, slow down, and speed up the video.

5.2 Procedure and Task

Participants, which we refer to as attackers, were introduced to the threat models and the input methods through explainer videos. We added control questions after each introduction where they had to guess a single-digit entry based on a given picture of the authentication scheme and the VR avatar's interaction to ensure they understand how input using each method works. We navigated them back to the explainer videos if they did not pass the control questions. Each block then displayed 8 PINs, 4 for each threat model. No participant attacked the same PIN more than once. Participants could provide up to three guesses and rate their confidence in their guess using a 5-point Likert scale. We concluded with a questionnaire and a file upload where participants could upload any notes they took.

5.3 Results

We recruited 22 participants (10 female, 12 male). They were on average 25.82 years (range: 19 - 45, SD=7.77). Compared to the real-world study of CueAuth, our sample is slightly more diverse (45.45% female participants vs 18.18% female participants, self-reported). Participants' age in both studies is almost identical with an average age of 25.82 years in our replication study and 26.9 in the original real-world study [52]. We analysed their successful attack rate, the Levenshtein distance between users' correct and attackers' closest guess, attackers' attack duration, and their level of confidence when performing observation attacks. We excluded the data of 6 participants (vs two in the real-world study [52]) as it was clear they did not put reasonable effort in the attacks (0 out of 24 attacks were successful). We discuss this decision further in Section 6.2.

Table 1: Single attacks against mid-air and eye gaze resulted in 0 successful attacks. Only 10.95% were successful on touch. Although repeated-video attacks were equally successful for touch and mid-air (59.38%), the Levenshtein distance shows that attacks on touch were closer to the correct PIN. Participants' level of confidence (1=not confident at all, 5=very confident) remained the same across the threat models in eye gaze, but repeated-video attacks significantly increased attackers' confidence in touch and mid-air.

	Single attack				Repeated-video attack			
	Success	**Distance**	**Confidence**	**Duration**	**Success**	**Distance**	**Confidence**	**Duration**
Touch	10.95%	2.03	1.74	103.00 s	59.38%	0.59	4.34	150.98 s
Mid-air	0.00%	2.94	1.33	79.68 s	59.38%	0.83	4.53	138.80 s
Eye Gaze	0.00%	3.55	1.06	63.48 s	0.00%	3.45	1.09	150.54 s

5.3.1 Successful Attack Rate. In the single attack threat model, only a few attacks (10.95%, SD=18.19%) on touch were successful. Not a single attack was successful against mid-air and eye gaze. When attackers were able to rewind the videos, slightly more than half of the attacks on touch (59.38%, SD=22.12%) and mid-air (59.38%, SD=27.2%) were successful, but not a single one on eye gaze. We found a significant effect of the threat model, $F_{1,15} = 125.952$, $p<0.05$, and the input method, $F_{1.340,20.097} = 37.426$, $p<0.05$, on attackers' attack rate. We also found a significant interaction effect (threat model × input method), $F_{2,30} = 30.829$, $p<0.05$. While follow-up analysis revealed a significant difference of the attack rates between the input methods in the single threat model, $F_{2,30} = 5.787$, $p<0.05$, post-hoc pairwise comparisons did not confirm these differences. In the repeated-video attack threat model, attacks on touch (59.38%) and mid-air (59.38%) were significantly more successful ($p<0.05$) than on eye gaze (0%). No other pairs were significant different. The same pairs were significantly different in the original CueAuth paper [52]. When comparing attackers' attack rate between the study types, we found no statistically significant three-way interaction effect (study type × input method × threat model), $F_{2,68} = 2.065$, $p = 0.135$, and no statistically significant two-way interaction (study type × input method), $F_{2,68} = 2.710$, $p = 0.074$; (study type × threat model), $F_{1,34} = 0.113$, $p = 0.738$.

Inline with CueAuth's real-world evaluation [52], our analysis revealed that repeated-video attacks on a VR avatar when providing input with touch and mid-air were more successful than on eye gaze. No differences between the input methods were found for single attacks in both our replication study and in the real-world study. However, it should be noted that single attacks on touch were to 10.9% successful in our replication study, while in the real-world study they were not successful at all (0%) [52]. The other was found in repeated-video attacks on touch with 59.38% successful attacks in our replication study and 74% in the real-world study [52]. The values of mid-air and eye gaze in our replication study match more accurately with the real-world study results [52]. Results are summarised in Table 1 and Table 2.

Observation #6 [Attack Rate]: Successful attack rates against VR avatars are largely similar to attacks against real users as done the original study. There is no evidence that attackers performed better when attacking a real-world user compared to a virtual avatar, and vice versa.

5.3.2 Levenshtein Distance. As done in CueAuth [52], we calculated the Levenshtein distances [62] between the correct PIN and the attacker's closest guess to it. We found a significant effect of the threat model, $F_{1,15} = 88.679$, $p<0.05$, and input method, $F_{2,30} = 170.284$, $p<0.05$, on the Levenshtein distance. We also found a significant interaction effect (threat model × input method), $F_{2,30} = 46.126$, $p<0.05$. Follow-up analysis revealed a significant effect of input method on the Levenshtein distance in case of single attacks, $F_{2,30} = 30.551$, $p<0.05$. Post-hoc pairwise comparisons revealed significant differences between all three input methods ($p<0.05$). Between touch (M=2.03, SD=1.07) and eye gaze (M=3.55, SD=0.56), touch (M=2.03, SD=1.07) and mid-air (M=2.94, SD=1.03), and eye gaze (M=3.55, SD=0.56) and mid-air (M=2.94, SD=1.03). We also found a significant effect of input method on the Levenshtein distance in case of repeated-video attacks, $F_{2,30} = 335.889$, $p<0.05$. Post-hoc pairwise comparisons revealed a significant difference between touch (M=0.59, SD=0.63) and eye gaze (M=3.45, SD=0.71), and mid-air (M=0.83, SD=0.86) and eye gaze (M=3.45, SD=0.71). No significant difference was found between touch and mid-air. Repeated-video attacks did not improve users' attacks on eye gaze, but when attacking touch and mid-air input ($p<0.05$). When comparing the Levenshtein distances of attackers' guesses to the original real-world study, we found a statistically significant three-way interaction effect (study type × input method × threat model), $F_{2,68} = 11.122$, $p<0.05$ and a statistically significant two-way interaction (method × study type), $F_{2,68} = 15.621$, $p<0.05$. Single-attacks on eye gaze in the real-world study were statistically significant closer ($p<0.05$) to the correct PINs (M=2.81, SD=0.76 vs M=3.55, SD=0.32), whereas single-attacks on touch input were closer ($p<0.05$) to the correct PINs in our replication study (M=2.03,SD=0.77 vs M=2.83,SD=0.67). In repeated-video attacks, guesses on eye gaze where statistically significant closer to the correct PINs in the real-world study (M=2.63, SD=0.72 vs M=3.45, SD=0.31). No other significant differences were found.

The differences within our replication study in the single-atack threat model are not inline with the real-world study. However, in repeated-video attacks, attacks on touch and mid-air were both significantly closer to the correct PINs than on eye gaze, which is inline with the real-world study. Both our VR and the real-world study [52] suggest that repeated-video attacks improved attackers' performance in all three input methods.

Observation #7 [Levenshtein Distance]: Attacks on mid-air were equally close in both study types, whereas attacks on eye gaze were more accurate in the real-world study. Attackers in the

replication study performed significantly closer single-attacks on touch input compared to the single-attacks in the real-world study. In both the real-world study and the replication study, repeated-video attacks were more successful on touch and mid-air than on eye gaze.

5.3.3 Attack Duration. When assessing attackers' attack duration, we found no effect of the input method on attack duration $F_{1.241,16.130} = 0.115$, $p = 0.792$. The values for touch, mid-air, and eye gaze are M=150.98 (SD=65.83), M=138.80 (SD=44.68), and M=150.54 (SD=33.19). When comparing the attack durations to the original real-world study, we found no statistically significant interaction effect (study type × input method) on attack duration, $F_{1.296,44.049} = 1.938$, $p = 0.168$. In both study types, attacks on mid-air were fastest. However, in the real-world study attacks on mid-air were also significantly faster than eye gaze. This was not the case in our replication study. All other pairs match – there were no significant differences between touch and mid-air, and touch and eye gaze in the real-world study of CueAuth [52] and in our replication study.

Observation #8 [Attack Duration]: There is no evidence that attackers in the replication study spent more or less time on their attacks than in the original real-world study. In both the replication study and the real-world study, we did not find a significant difference of attack duration between touch and mid-air, and touch and eye gaze. However, attacks on touch and mid-air in our replication study were notably longer than in the real-world study [52] (touch: 150.98 s vs 103.9 s [52]; mid-air: 138.80 s vs 91.9 s [52]).

5.3.4 Attackers' Confidence. When assessing attackers' level of confidence, we found a significant effect of the threat model, $F_{1,15} = 252.842$, $p<0.05$, and input method, $F_{2,30} = 284.938$, $p<0.05$. We also found a significant interaction effect (threat model × input method), $F_{2,30} = 147.413$, $p<0.05$. Follow-up analysis revealed a significant main effect of input method on attackers' confidence when performing single attacks, $F_{2,30} = 15.838$, $p<0.05$, and repeated-video attacks, $F_{2,30} = 341.548$, $p<0.05$. In terms of single attacks, we found significant differences ($p<0.05$) of attackers' confidence between touch (M=1.74, SD=0.48) and mid-air (M=1.33, SD=0.35), and touch (M=1.74, SD=0.48) and eye gaze (M=1.06, SD=0.14). For repeated-video attacks, attackers were significantly less confident about their guesses when PINs were entered with eye gaze (M=1.09, SD=0.27) compared to touch (M=4.34, SD=0.74) and mid-air (M=4.53, SD=0.54). We also found that attackers' confidence was significantly higher ($p<0.05$) in repeated-video attacks in case of touch (M=4.34, SD=0.74) and mid-air (M=4.53, SD=0.54) compared to single attacks on touch (M=1.74, SD=0.48) and mid-air (M=1.33, SD=0.35). No other pairs were significant. Results are summarised in Table 1. Attackers' confidence together with their notably low successful attack rate on eye gaze is inline with prior work that emphasises the high resistance to observations of gaze-based authentication [28, 46, 61, 68]. When comparing attackers' confidence between the two study types, we found a significant three-way interaction effect (input method × threat model × study type), $F_{1.399,47.582} = 10.485$, $p<0.05$. We also found a statistically significant two-way interaction effect (input method × study type), $F_{1.620,55.076} = 15.403$, $p<0.05$. Follow-up analysis revealed a significant difference ($p<0.05$) between attackers' confidence in the real-world and replication study when attacking mid-air and eye gaze in both threat models. The values for mid-air in single attacks are M=1.33 (SD=0.35) in VR and M=2.03 (SD=1.01) in the real world. For eye gaze the values are M=1.06 (SD=0.14) in VR and M=1.92 (SD=1.04) in the real world. For repeated-video attacks the values for mid-air are M=4.53 (SD=0.54) in VR and M=3.71 (SD=1.16) in the real world. The values for eye gaze are M=1.09 (SD=0.27) in VR and M=2.28 (SD=1.15) in the real world. We did not find any other statistically significant differences. Results are summarised in Table 2. Inline with the real-world study, we found that attackers in repeated-video attacks on touch and mid-air were more confident than in single attacks. Attackers in repeated-video attacks on touch and mid-air were also significantly more confident than in repeated-video attacks on eye gaze — this was the case for both the real-world study and our replication study.

Observation #9 [Attackers' Confidence]: There is no evidence that attackers were more confident in attacking touch in either one of the study types. However, attackers were more confident in their single-view attacks on mid-air and eye gaze in the real-world study. They were also more confident when performing repeated-video attacks on eye gaze in the real world, but less confident when attacking mid-air.

5.3.5 Qualitative Feedback. Unlike data collected through semi-structured interviews in the real-world study [52], we relied on open questions at the end of our online survey. We pre-defined three main areas of interest as also done in the real-world study [52]: attackers' attacking strategy and their security and usability perception when using the input methods.

Attacking Strategy: Attackers' attacking strategies in our replication study match those used in the original real-world study [52] to a great extent. While the majority mostly noted down the PIN numbers on a piece of paper, others drew figures or even drew a sketch of the virtual environment. Participants raised that further training could help them in running successful attacks. Single-view attacks on touch and mid-air were perceived as too fast. Participants found it challenging to switch between the hand movements and the digits on the screen. Attacks on eye gaze were perceived as too hard. Participants mentioned they could hardly see the eyes move. Instead, they guessed the direction of users' eyes to indicate on which digit the user gazes at. Others mentioned that the combination of focusing on users' eye movements and the screen put them off. In repeated-video attacks, attackers made use of rewinding and slowing down the videos. This was mentioned frequently in touch and mid-air, but not in eye gaze. In mid-air, attackers mentioned that observing one-handed interactions was easier. Similar to single attacks on eye gaze, attackers perceived eye gaze as "hard" to attack and mentioned that slowing down the videos did not help recognise eye movements. Some attackers reported having a vague idea of the entered digits, but could not use these information to provide successful attacks on eye gaze. This is inline with the successful attack rate (Observation #1 and Section 5.3.1).

Security Perception: Participants found eye gaze the most secure input method, which is inline with our findings in the in-VR

usability study and the original real-world study [52]. Similar to the real-world study, opinions differed when asking about the least secure method: 11 found touch the least secure input method, while 5 found mid-air the least secure input method.

Usability Perception: The feedback we received from the participants in the security study aligned with the findings in the in-VR usability study and the real-world study [52]. Touch was defined as easy to use, convenient, and most practical. P4 and P6 mentioned that entering a PIN can be done discreetly by covering up the hands. On the other hand, P16 mentioned that the on-screen gestures could linger on the screen after input. Others mentioned touch input is easy to attack and that people are already used to attack this input method. Other comments were about the hygiene of the method. For example, P4 raised the concern about the infection risk in a post-pandemic world where you have to touch surfaces. Participants mentioned that providing input with mid-air feels a bit like being a fool in public and looks strange. Overall, they were reserved towards the social acceptability of mid-air. For example, P11 and P12 stated that mid-air reminds them of playing games on a Nintendo WII or in VR. P14 mentioned that using mid-air feels like sharing the PIN with everyone around and P12 mentioned that using such mid-air gestures over-complicates input a lot. Participants were reserved towards the usability of eye gaze. Entering PINs with eye gaze was considered to be hard to use, hard to learn, and long-winded. P5 mentioned that the concept feels weird. There were many concerns regarding the accuracy of such a system. For example, participants mentioned eye gaze could lead to many errors that could prompt users to go through the authentication process multiple times. On the positive side, P4 mentioned that eye gaze could support disabled people when interacting with such systems.

Observation #10 [Qualitative Feedback]: Many of the voiced comments match those voiced in our in-VR usability study and in the original real-world study — indicating that users' perception of the methods in terms of security and usability remained the same across the studies and was not influenced by the virtual avatar and study type.

6 DISCUSSION & FUTURE WORK

Users' perceptions of the usability and security of the input methods collected through our studies match to a great extent the perceptions of the real-world study participants (Observation #4 and #10). Our validation also revealed many similarities between the quantitative measures of our VR and the real-world studies. Results are summarised in Table 2. Similar to the design implications in the real-world study [52] we can deduce the following implications from our experiments:

- **Design Implication 1:** Eye gaze is the most secure input method, but the slowest (see Table 2). This suggests that eye gaze is suitable when authentication frequency is low and subtle authentication is required, which is inline with the design implication 1 reported in the real-world study [52].
- **Design Implication 2:** When comparing eye gaze with mid-air in our in-VR usability and online security study, we also conclude that mid-air is more usable than eye gaze, but eye gaze is more secure (see also design implication 2 in [52]).
- **Design Implication 3:** Our qualitative results suggest mid-air is not suitable for public spaces (e.g., requires additional space and *"looks like a jumping jack", P17*); thus users should be able to opt for alternative modalities. This finding is also depicted in [52]'s design implication 3.

While we can deduce the same design implications as found in the original real-world study [52], there are also measures that do not match and require further discussions. We discuss the suitability of using VR to evaluate authentication prototypes in the light of our findings. We first discuss similarities and differences between our conceptual VR replication and CueAuth's usability findings [52]. We then discuss the impact of using VR avatars in observation resistance studies by drawing upon our security study results and conclude with potential research directions.

6.1 In-VR Usability Evaluation: Users' Performance and Perception

We found similar entry accuracies in all three input methods (Observation #1) and there is no evidence that users' perceived workload was significantly different in our VR study when providing input with touch and mid-air (Observation #3). However, building upon the work by De Luca et al. [18] and Knierim et al. [57], our results suggest that tracking accuracy, in our case hand and eye tracking, can have a notable impact on a system's usability. This is apparent in our study as follows: Compared to the real-world usability study of CueAuth [52], input using touch was significantly slower in VR, while input using gaze was significantly faster (Observation #2). Eye gaze was perceived as significantly less physically demanding, less frustrating, and required less effort in VR (Observation #3), and was the most preferred input method in our VR study, whereas it was the least preferred one in the real-world study (Observation #5). To summarise, these differences between the real-world and our VR study show that VR studies do not necessarily provide the often desired "all-in-one solution" [112] and researchers need to have a clear vision what they expect from such evaluations as not all measures may be transferable to the real world; yet, they have lots of potential for researchers to evaluate and deploy their prototype systems. We discuss these findings below.

6.1.1 Users' Virtual Hands: Contrary to mid-air and eye gaze where no additional virtual artefacts are required when providing input, touch requires visualising users' hands because they leverage the visual feedback of their hands in the real-world to provide precise input. The virtual hand then lies between the user and the interaction. This can affect users' performance; our participants voiced that the virtual hand did not always map with their real hand. Although participants associated touch input with many positive attributes in our VR study, their entries were not as fast as those logged in the real-world study. This confirms and extends the findings by Knierim et al. [57] who used OptiTrack cameras for finger tracking while typing in VR. Their results also showed that users' typing performance is affected by the avatar hands [57]. We intentionally abstained from using an OptiTrack or other high-end sensors because one of the goals of VR studies is to cut down prototyping

Table 2: Our in-VR study achieved similar results in terms of users' entry accuracy and perceived workload compared to the real-world study. There is no evidence that attackers' attack rate and attack duration differ significantly between the study types. However, measures such as entry time, the distance of attackers' guesses to the correct PINs and their confidence were significantly different between the two study types. ⋆ The reported values are for both single and repeated-video attacks with the notation *single | repeated.*

Measures	Real World [52]	VR	Measures	Real World [52]	Observing Virtual Avatars
Entry Accuracy			**Attack Rate⋆**		
Touch	93.38%	89.97%	Touch	0.00% | 74%	10.95% | 59.38%
Mid-Air	84.19%	80.42%	Mid-Air	0.01% | 64%	0.00% | 59.38%
Eye Gaze	82.72%	83.75%	Eye Gaze	0.03% | 0.05%	0.00% | 0.00%
Entry Time			**Levenshtein Distance⋆**		
Touch	3.73 s	6.06 s	Touch	2.83 | 0.50	2.03 | 0.59
Mid-Air	5.51 s	5.54 s	Mid-Air	2.78 | 0.64	2.94 | 0.83
Eye Gaze	26.35 s	16.75 s	Eye Gaze	2.81 | 2.63	3.55 | 3.45
Perceived Workload			**Attack Duration⋆**		
Touch	23.25	34.83	Touch	N/A | 103.9 s	103.00 s | 150.98 s
Mid-Air	39.584	34	Mid-Air	N/A | 91.9 s	79.68 s | 138.80 s
Eye Gaze	46.54	30.33	Eye Gaze	N/A | 163.4 s	63.48 s | 150.54 s
Preferred Input Method			**Attacker Confidence⋆**		
Touch	56	37	Touch	1.89 | 3.75	1.74 | 4.34
Mid-Air	33	41	Mid-Air	2.02 | 3.71	1.33 | 4.53
Eye Gaze	31	45	Eye Gaze	1.92 | 2.28	1.06 | 1.09

expenses. Although such technological limitations may disappear due to improved finger tracking and acquaintance of users with VR, they also suggest that VR may sometimes not be able to provide users with exactly the same experience they would face in reality (e.g., [89]).

6.1.2 The Consequences of Improved Eye Tracking Systems: Contrary to the real-world study, participants in our VR study preferred eye gaze over the other methods (Observation #5) and provided faster entries than in the original real-world study (Observation #2). This difference in performance and perception of gaze is due to the different eye tracking systems used. In the real-world study, Khamis et al. [52] used a stationary eye tracker mounted at the bottom of the display. This likely resulted in the eye tracking quality being influenced by the user's height, distance to the display, and the ambient and lighting conditions in the room. These are known problems in studies that involve stationary eye trackers [17, 51, 76]. In fact, the authors of CueAuth admit that "error rates and entry times are influenced by [their] setup and implementation" [52]. In our case, we used an eye tracker that is integrated in the head-mounted display. This meant that many of the artefacts that typically reduce eye tracking quality were absent in our VR study. For example, eye tracking in our study was independent of the user's position, their height, and the surrounding lighting conditions. This explains why users were faster and preferred using eye gaze in the VR study. However, it is important to note that our findings were also dependent on the technology used in our replication study and in the original real-world evaluation of CueAuth [52]; a more advanced stationary eye tracker than the one used in the original real-world study [52] may achieve results similar to those found in our replication study. Our findings also show that VR studies could help mitigate limitations of hardware used in the real world, but could also be misleading. For example, if a researcher wants to assess the usability of gaze-based authentication assuming ideal tracking conditions, then VR would help them achieve that. On the downside, if a researcher wants to assess the usability of the same scheme with noise and other external factors put into consideration, then they would need to account for these factors in VR, or otherwise risk being mislead into thinking that the scheme works better than it actually does.

There are similar trade-offs that HCI and USEC researchers have traditionally considered when comparing lab and field studies, with one optimising for control while the other optimising for ecological validity. What is important to note here is that **transferability of quantitative results from VR to the real world highly depends on how well reality and its limitations are emulated**. It is also important to recognise that **VR studies are not an alternative to lab or field studies**, but rather can complement them by enabling large-scale evaluations and cost-efficient prototyping.

6.2 Attackers' Performance in Online Security Studies and the Use of Virtual Avatars

Voit et al. [104] observed that participants in their online method were less engaged than in studies where a researcher was present. The poor performance of six participants in our security study suggests that this phenomenon (see also [16]) was also present in our security study. Data quality is one of the major concerns when using online crowd-sourcing platforms [78]. In our security study, we recruited the same number of participants (N=22) as in the real-world security study of CueAuth [52]. While in the original work the authors excluded two participants due to their poor performance (0 out of 24 attacks were successful), we excluded six participants due to the same reason. We could find repetitive

guesses throughout the survey (e.g., "1234") and repeatedly wrongly answered control questions, indicating that some participants did not participate meaningfully in the study. This suggests that the 22 participants in the real-world CueAuth study [52] felt more committed to their participation than in our online study, which extends the findings by Clifford et al. [13] and Voit et al. [104] who found significant differences in distractions between online and lab studies [13] and that online surveys receive significantly more low-quality responses (e.g., lower word counts) than real-world studies [104]. Our online security study emphasises the importance of filtering out these low-quality responses in security studies and extends the findings by Redmiles et al. [85] and Fahl et al. [24] who argue that the ecological validity can be improved by filtering out such cases. After excluding the six participants, our security study results indeed match with the real-world evaluation to a great extent. There is no evidence that the successful attack rates differ significantly between the two study types (Observation #6), the performance of attacks against the different input methods followed the same pattern across both studies, and significant differences were found between the same pairs in VR and the real-world. While more general measures such as attackers' successful attack rate (Observation #6) match between the study types, more specific measures such as the Levenshtein distance differ (Observation #7).

To summarise, **researchers should be aware of low-quality responses when conducting security evaluations online**, but **a well-defined exclusion criteria contributes to receiving research findings that are highly transferable to the real world.** The results also show that **observing virtual avatars during authentication reveals insights about how resistant the used scheme is against observations**. The idea of using virtual avatars instead of humans in evaluations raises interesting directions for future work. For example, avatars can be programmed to mimic movements of users with disabilities or conditions that impact their mobility e.g., Parkinson's disease, or users that are challenging to recruit e.g., children. The use of avatars to simulate these user groups makes VR studies promising in cases where recruitment is challenging due to ethical or logistic reasons; yet, it is researchers' responsibility to act ethically and morally [10, 97].

6.3 VR Studies for Human-centred Usability, Privacy, and Security Research

In this work, we focused on authentication as a sub-domain of USEC research. However, USEC research covers much more including privacy on IoT devices [65, 93, 113], users' perception of technologies (e.g., drones [12]) and how they affect privacy and security, or social engineering [36, 59]. Based on the results and our experience conducting this work, we discuss two potential research directions where we see great potential of VR studies as a study method. Our work lays the foundation for future user studies exploring the potential of VR studies for usable privacy and security research. Follow-up research should consider replication studies similar to our work to draw further insights about the strengths and weaknesses of VR studies.

6.3.1 Virtual Field Studies: Mäkelä et al. [64], Savino et al. [89], and Voit et al. [104] have taken the first steps towards using VR as a research method for virtual human-centred studies. A promising next step could be a combination of virtual field studies [64, 89], virtual artefacts [104], and authentication as studied in our work. How can we incorporate all the benefits of field studies into human-centred evaluations while eliminating their additional costs and complexity? There are many authentication schemes (e.g., [19, 53]) that were studied without the potential effect of external factors such as bystanders or vivid contexts. How does user's behaviour translate to the real world when interacting with authentication systems in a more vivid environment in VR? While such virtual field studies would not replace real-world field studies, they have the benefit of (a) decreasing the additional costs researchers face when going "into the wild" (e.g., costs of hardware prototyping, booking access to specific locations); (b) eliminating confounding variables that are hard to pinpoint in real-world field studies such as the impact of bystanders or different lighting conditions; and (c) they may enable researchers to conduct research in contexts that are otherwise challenging due to ethical or legal constraints.

One concrete application could be to study novel authentication systems at ATMs and the impact of bystanders on corresponding user authentications. Do users provide faster entries, but make more errors when they feel being observed by a virtual character? VR could also transform physical labs into smart home environments that can enable researchers to investigate user behaviour in more ecologically valid settings. For example, when interacting with IoT devices to study users' awareness [94] or to study users' perception of different authentication mechanisms at doors [71]. Using VR studies for such evaluations can be particularly valuable as they can immerse users into different contexts. Participants in lab studies often manifest *"demand characteristics"* where they subconsciously change their behaviour to fit the experimenter's purpose [21, 86] (Hawthorne effect [2]). In human-centred privacy and security research, researchers often rely on cameras to record user interactions for follow-up evaluations (e.g., [7, 8, 18]). In such cases, participants are often aware of the recording devices and the presence of the experimenter who often shares the same physical space. In contrast, one of our participants, P13, voiced that she was totally unaware of the experimenter during our in-VR usability study. VR studies of that type could shed more light on the Hawthorne effect [2] within and beyond USEC research.

6.3.2 Remote VR Studies: Another promising research direction is to conduct fully remote VR studies. For example, the hardware used in our studies has become more accessible and affordable in the past (e.g., the HTC VIVE Pro Eye [42]) — making it more likely to encounter such devices in many households in the near future. While we conducted only our security study online, moving the usability study to a fully remote VR experience would not require much additional effort. As done in our in-VR usability study, questionnaires can be integrated into the VR experience without much effort [4, 81], material such as the participant information sheet can be distributed in advance of the study, and semi-structured interviews can be done via videotelephony. Remote VR studies could enable large-scale human-centred usability, privacy, and security evaluations of replications of physical prototypes.

This approach is deemed promising beyond USEC research; for example, recent suggestions in running user studies that involve virtual and augmented reality amidst COVID-19 include collaborations

across labs to provide participants for each other's experiments, and building an infrastructure that provides equipment to a pool of participants [95].

7 LIMITATIONS

It is important to consider the following technological and experimental design limitations when interpreting our findings. First, we studied and replicated only one real-world authentication scheme (CueAuth [52]) and showed through two user studies that VR can serve as a suitable test-bed for the usability and security evaluation. Although we studied the usability and security of a breadth of input methods that are frequently used in security research: touch [19, 107], mid-air [1, 6], and eye gaze [20, 46, 61], it is important to note that substantial future research (e.g., additional replication studies, replications using participants' own VR headsets) is necessary to be able to generalise our results to a larger set of authentication systems and other usable privacy and security systems. Second, the security evaluation is based on two specific threat models: 1) single observation attacks and 2) repeated video attacks. In both the replication study and the real-world study [52] the attacks were single-person attacks through optimal views on the authentication scheme and the interaction – this depicts a best case scenario for the attacker. Using non-optimal or user-defined views, or more advanced threat models (e.g., multiple observers [48]), could result in different findings. Third, following Khamis et al.'s [52] study design means facing the same limitations. CueAuth's evaluation was dependent on the used hardware in both the real-world and our VR study. Other hardware such as OptiTrack systems could have lead to more accurate tracking. However, if (remote) VR studies are to become mainstream, they have to leverage personal commodity hardware that a typical VR user would have. It is unlikely that average VR users will own a high-end tracking systems like OptiTrack. Finally, replication studies are generally challenging [109]; the largest replication study to date attempted to replicate 100 studies and succeeded only in 39% of them [14]. Human test subjects consciously or sub-consciously remember previous experiences that can impact their thoughts, behaviour, and performance; thus, experiments can result in different results due to the non-uniformity of nature [35, 90]. While the original work was published in 2018 [52], our studies were conducted two years later during a pandemic [77] — which has to be noted.

8 CONCLUSION

Currently used empirical methods for evaluating physical authentication systems are not suitable for large-scale usable privacy and security research that involves non-commodity hardware. Deploying physical authentication prototypes to, for example, evaluate a system's usability and security and understand users' behaviour during interactions makes studies of that type often costly and time consuming. Given this motivation, we conducted a conceptual replication study of a real-world authentication system, CueAuth [52], and investigated the suitability of Virtual Reality (VR) studies to evaluate the usability and security of real-world authentication systems. We report on results from a lab-based in-VR usability study (N=20) and an online security study through VR recordings (N=22). Our lab-based in-VR usability study suggests that many usability measures of VR studies are transferable to the real world. For example, users achieved similar entry accuracies, reported similar perceived workload when authenticating with touch, mid-air, and eye gaze, and shared a similar security perception of the input methods across our VR study and the original real-world study [52]. However, notable longer entries in VR were found when using touch — indicating that introducing virtual artefacts, through, for example, finger tracking can have a negative impact on users' performance while wearable eye trackers instead of static eye trackers can have a positive impact on users' performance. We observed great similarities between our security study where attackers performed observations on a virtual avatar compared to observations on a human in the real world [52]. Attack rates do not significantly differ, and there is no evidence that attackers spent more or less time on their attacks in our online study compared to the real-world study. However, measures like the accuracy of attackers' guesses and their confidence differed significantly between the two study types, with users in the real-world study being closer to correct guesses on eye gaze, but not when performing single-view attacks on touch. Through this work, we provide insights into the potential, strengths, and current weaknesses of using VR for holistic usability and security evaluations of real-world authentication schemes. Our results suggest studies of that type can have a number of advantages over traditional evaluation methods such as lab, online, or field studies, but studies of that nature also come with limitations that need to be kept in mind.

Through our investigation of using VR as a test-bed for real-world authentication schemes we hope to open the door for follow-up research to establish VR studies as a fundamental evaluation paradigm in research on authentication prototypes and related USEC research domains.

ACKNOWLEDGEMENTS

We thank all participants for taking part in our user studies. We also thank the external reviewers and our ACs whose comments significantly improved the paper. This publication was supported by the University of Edinburgh and the University of Glasgow jointly funded PhD studentships. This work was also partially supported by the Royal Society of Edinburgh (award number #65040) and an EPSRC New Investigator Award (EP/V008870/1).

REFERENCES

[1] Yasmeen Abdrabou, Mohamed Khamis, Rana Mohamed Eisa, Sherif Ismail, and Amrl Elmougy. 2019. Just Gaze and Wave: Exploring the Use of Gaze and Gestures for Shoulder-Surfing Resilient Authentication. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications* (Denver, Colorado) *(ETRA '19)*. Association for Computing Machinery, New York, NY, USA, Article 29, 10 pages. https://doi.org/10.1145/3314111.3319837

[2] John G Adair. 1984. The Hawthorne effect: a reconsideration of the methodological artifact. *Journal of applied psychology* 69, 2 (1984), 334. https://doi.org/10.1037/0021-9010.69.2.334

[3] Sara Albakry, Kami Vaniea, and Maria K. Wolters. 2020. What is This URL's Destination? Empirical Evaluation of Users' URL Reading. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376168

[4] Dmitry Alexandrovsky, Susanne Putze, Michael Bonfert, Sebastian Höffner, Pitt Michelmann, Dirk Wenig, Rainer Malaka, and Jan David Smeddinck. 2020. Examining Design Choices of Questionnaires in VR User Studies. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–21. https://doi.org/10.1145/3313831.3376260

[5] K. Althobaiti, G. Rummani, and K. Vaniea. 2019. A Review of Human- and Computer-Facing URL Phishing Features. In *2019 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. EuroS&PW, Stockholm, Sweden, 182–191. https://doi.org/10.1109/EuroSPW.2019.00027

[6] Ilhan Aslan, Andreas Uhl, Alexander Meschtscherjakov, and Manfred Tscheligi. 2014. Mid-Air Authentication Gestures: An Exploration of Authentication Based on Palm and Finger Motions. In *Proceedings of the 16th International Conference on Multimodal Interaction* (Istanbul, Turkey) *(ICMI '14)*. Association for Computing Machinery, New York, NY, USA, 311–318. https://doi.org/10.1145/2663204.2663246

[7] Adam J. Aviv, John T. Davin, Flynn Wolf, and Ravi Kuber. 2017. Towards Baselines for Shoulder Surfing on Mobile Authentication. In *Proceedings of the 33rd Annual Computer Security Applications Conference* (Orlando, FL, USA) *(ACSAC 2017)*. Association for Computing Machinery, New York, NY, USA, 486–498. https://doi.org/10.1145/3134600.3134609

[8] Adam J. Aviv, Flynn Wolf, and Ravi Kuber. 2018. Comparing Video Based Shoulder Surfing with Live Simulation. In *Proceedings of the 34th Annual Computer Security Applications Conference* (San Juan, PR, USA) *(ACSAC '18)*. Association for Computing Machinery, New York, NY, USA, 453–466. https://doi.org/10.1145/3274694.3274702

[9] Louise Barkhuus and Jennifer A. Rode. 2007. From Mice to Men - 24 Years of Evaluation in CHI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '07)*. Association for Computing Machinery, New York, NY, USA, 1. https://doi.org/10.1145/1240624.2180963

[10] Philip AE Brey. 2011. Anticipatory technology ethics for emerging IT. *CEPE 2011: Crossing Boundaries* 6 (2011), 13.

[11] Kelly Caine. 2016. Local Standards for Sample Size at CHI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 981–992. https://doi.org/10.1145/2858036.2858498

[12] Victoria Chang, Pramod Chundury, and Marshini Chetty. 2017. Spiders in the Sky: User Perceptions of Drones, Privacy, and Security. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 6765–6776. https://doi.org/10.1145/3025453.3025632

[13] Scott Clifford and Jennifer Jerit. 2014. Is There a Cost to Convenience? An Experimental Comparison of Data Quality in Laboratory and Online Studies. *Journal of Experimental Political Science* 1, 2 (2014), 120–131. https://doi.org/10.1017/xps.2014.5

[14] Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349, 6251 (2015), 13. https://doi.org/10.1126/science.aac4716

[15] Jessica Colnago, Summer Devlin, Maggie Oates, Chelse Swoopes, Lujo Bauer, Lorrie Cranor, and Nicolas Christin. 2018. "It's Not Actually That Horrible": Exploring Adoption of Two-Factor Authentication at a University. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3173574.3174030

[16] Frédéric Dandurand, Thomas R Shultz, and Kristine H Onishi. 2008. Comparing online and lab methods in a problem-solving experiment. *Behavior research methods* 40, 2 (2008), 428–434. https://doi.org/10.3758/BRM.40.2.428

[17] Beymer David and Flickner Myron. 2003. Eye gaze tracking using an active stereo head. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, Vol. 2. IEEE Computer Society, USA, II–451. https://doi.org/10.1109/CVPR.2003.1211502

[18] Alexander De Luca, Marian Harbach, Emanuel von Zezschwitz, Max-Emanuel Maurer, Bernhard Ewald Slawik, Heinrich Hussmann, and Matthew Smith. 2014. Now You See Me, Now You Don't: Protecting Smartphone Authentication from Shoulder Surfers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 2937–2946. https://doi.org/10.1145/2556288.2557097

[19] Alexander De Luca, Emanuel von Zezschwitz, Ngo Dieu Huong Nguyen, Max-Emanuel Maurer, Elisa Rubegni, Marcello Paolo Scipioni, and Marc Langheinrich. 2013. Back-of-Device Authentication on Smartphones. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. Association for Computing Machinery, New York, NY, USA, 2389–2398. https://doi.org/10.1145/2470654.2481330

[20] Alexander De Luca, Roman Weiss, and Heiko Drewes. 2007. Evaluation of Eye-Gaze Interaction Methods for Security Enhanced PIN-Entry. In *Proceedings of the 19th Australasian Conference on Computer-Human Interaction: Entertaining User Interfaces* (Adelaide, Australia) *(OZCHI '07)*. Association for Computing Machinery, New York, NY, USA, 199–202. https://doi.org/10.1145/1324892.1324932

[21] Nicola Dell, Vidya Vaidyanathan, Indrani Medhi, Edward Cutrell, and William Thies. 2012. "Yours is Better!": Participant Response Bias in HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 1321–1330. https://doi.org/10.1145/2207676.2208589

[22] Statista Research Department. 2019. *Share of respondents who own a virtual reality headset computing device in the United Kingdom in 2019, by generation*. https://www.statista.com/statistics/1044070/uk-virtual-reality-headset-ownership/ accessed 15 September 2020.

[23] Alan Dix, Janet Finlay, Gregory D Abowd, and Russell Beale. 2003. *Human-computer interaction*. Pearson Education, USA.

[24] Sascha Fahl, Marian Harbach, Yasemin Acar, and Matthew Smith. 2013. On the Ecological Validity of a Password Study. In *Proceedings of the Ninth Symposium on Usable Privacy and Security* (Newcastle, United Kingdom) *(SOUPS '13)*. Association for Computing Machinery, New York, NY, USA, Article 13, 13 pages. https://doi.org/10.1145/2501604.2501617

[25] Timothy J. Forman, Daniel S. Roche, and Adam J. Aviv. 2020. Twice as Nice? A Preliminary Evaluation of Double Android Unlock Patterns. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3334480.3382922

[26] Alexander Gamero-Garrido, Stefan Savage, Kirill Levchenko, and Alex C. Snoeren. 2017. Quantifying the Pressure of Legal Risks on Third-Party Vulnerability Research. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (Dallas, Texas, USA) *(CCS '17)*. Association for Computing Machinery, New York, NY, USA, 1501–1513. https://doi.org/10.1145/3133956.3134047

[27] Simson Garfinkel and Heather Richter Lipford. 2014. Usable security: History, themes, and challenges. *Synthesis Lectures on Information Security, Privacy, and Trust* 5, 2 (2014), 1–124.

[28] Ceenu George, Daniel Buschek, Andrea Ngao, and Mohamed Khamis. 2020. GazeRoomLock: Using Gaze and Head-pose to Improve the Usability and ObservationResistance of 3D Passwords in Virtual Reality. In *Augmented Reality, Virtual Reality, and Computer Graphics*. Springer International Publishing. https://doi.org/10.1007/978-3-030-58465-8_5

[29] Ceenu George, Mohamed Khamis, Emanuel von Zezschwitz, Marinus Burger, Henri Schmidt, Florian Alt, and Heinrich Hussmann. 2017. Seamless and Secure VR: Adapting and Evaluating Established Authentication Systems for Virtual Reality. In *Network and Distributed System Security Symposium (NDSS 2017) (USEC '17)*. NDSS, 12. https://doi.org/10.14722/usec.2017.23028

[30] Yiannis Georgiou and Eleni A. Kyza. 2017. The development and validation of the ARI questionnaire: An instrument for measuring immersion in location-based augmented reality settings. *International Journal of Human-Computer Studies* 98 (2017), 24 – 37. https://doi.org/10.1016/j.ijhcs.2016.09.014

[31] Lisa M Given. 2008. *The Sage encyclopedia of qualitative research methods*. Sage publications. https://doi.org/10.4135/9781412963909.n381

[32] Gene V Glass, Percy D. Peckham, and James R. Sanders. 1972. Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analyses of Variance and Covariance. *Review of Educational Research* 42, 3 (1972), 237–288. https://doi.org/10.3102/00346543042003237

[33] Dominic Gorecky, Mohamed Khamis, and Katharina Mura. 2017. Introduction and establishment of virtual training in the factory of the future. *International Journal of Computer Integrated Manufacturing* 30, 1 (2017), 182–190. https://doi.org/10.1080/0951192X.2015.1067918

[34] Saul Greenberg and Bill Buxton. 2008. Usability Evaluation Considered Harmful (Some of the Time). In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. Association for Computing Machinery, New York, NY, USA, 111–120. https://doi.org/10.1145/1357054.1357074

[35] Stephan Guttinger. 2020. The limits of replicability. *European Journal for Philosophy of Science* 10, 2 (2020), 10. https://doi.org/10.1007/s13194-019-0269-1

[36] Christopher Hadnagy. 2010. *Social engineering: The art of human hacking*. John Wiley & Sons.

[37] Marian Harbach, Emanuel Von Zezschwitz, Andreas Fichtner, Alexander De Luca, and Matthew Smith. 2014. It's a Hard Lock Life: A Field Study of Smartphone (Un)Locking Behavior and Risk Perception. In *Proceedings of the Tenth USENIX Conference on Usable Privacy and Security* (Menlo Park, CA) *(SOUPS '14)*. USENIX Association, USA, 213–230.

[38] Sandra G Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. Sage publications Sage CA: Los Angeles, CA, 904–908. https://doi.org/10.1177/154193120605000909

[39] Marc Hassenzahl, Michael Burmester, and Franz Koller. 2003. AttrakDiff: A questionnaire to measure perceived hedonic and pragmatic quality. In *Mensch & Computer*, Vol. 57. 187–196.

[40] Eiji Hayashi, Oriana Riva, Karin Strauss, A. J. Bernheim Brush, and Stuart Schechter. 2012. Goldilocks and the Two Mobile Devices: Going beyond All-or-Nothing Access to a Device's Applications. In *Proceedings of the Eighth Symposium on Usable Privacy and Security* (Washington, D.C.) *(SOUPS '12)*. Association for Computing Machinery, New York, NY, USA, Article 2, 11 pages. https://doi.org/10.1145/2335356.2335359

[41] Jerry L. Hintze and Ray D. Nelson. 1998. Violin Plots: A Box Plot-Density Trace Synergism. *The American Statistician* 52, 2 (1998), 181–184. https://doi.org/10.1080/00031305.1998.10480559

[42] HTC. 2016. *HTC VIVE Pro Eye.* https://www.vive.com/uk/product/vive-pro-eye/ accessed 17 August 2020.

[43] HaptX Inc. 2020. *HaptX Gloves Development Kit.* https://haptx.com/ accessed 15 September 2020.

[44] Adam Jenkins, Pieris Kalligeros, Kami Vaniea, and Maria K Wolters. 2020. "Anyone Else Seeing this Error?": Community, System Administrators, and Patch Information. (2020), 15.

[45] Maurits Kaptein and Judy Robertson. 2012. Rethinking Statistical Analysis Methods for CHI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 1105–1114. https://doi.org/10.1145/2207676.2208557

[46] Christina Katsini, Yasmeen Abdrabou, George E. Raptis, Mohamed Khamis, and Florian Alt. 2020. The Role of Eye Gaze in Security and Privacy Applications: Survey and Future HCI Research Directions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–21. https://doi.org/10.1145/3313831.3376840

[47] Mohamed Khamis, Florian Alt, Mariam Hassib, Emanuel von Zezschwitz, Regina Hasholzner, and Andreas Bulling. 2016. GazeTouchPass: Multimodal Authentication Using Gaze and Touch on Mobile Devices. In *Proceedings of the 34th Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems* (San Jose, CA, USA) *(CHI EA '16)*. ACM, New York, NY, USA, 6. https://doi.org/10.1145/2851581.2892314

[48] Mohamed Khamis, Linda Bandelow, Stina Schick, Dario Casadevall, Andreas Bulling, and Florian Alt. 2017. They are all after you: Investigating the Viability of a Threat Model that involves Multiple Shoulder Surfers. In *Proceedings of the 16th International Conference on Mobile and Ubiquitous Multimedia* (Stuttgart, Germany) *(MUM '17)*. ACM, New York, NY, USA, 5. https://doi.org/10.1145/3152832.3152851

[49] Mohamed Khamis, Malin Eiband, Martin Zürn, and Heinrich Hussmann. 2018. EyeSpot: Leveraging Gaze to Protect Private Text Content on Mobile Devices from Shoulder Surfing. *Multimodal Technologies and Interaction* 2, 3, Article 45 (2018), 15 pages. https://doi.org/10.3390/mti2030045

[50] Mohamed Khamis, Mariam Hassib, Emanuel von Zezschwitz, Andreas Bulling, and Florian Alt. 2017. GazeTouchPIN: Protecting Sensitive Data on Mobile Devices using Secure Multimodal Authentication. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction* (Glasgow, Scotland) *(ICMI 2017)*. ACM, New York, NY, USA, 5. https://doi.org/10.1145/3136755.3136809

[51] Mohamed Khamis, Axel Hoesl, Alexander Klimczak, Martin Reiss, Florian Alt, and Andreas Bulling. 2017. EyeScout: Active Eye Tracking for Position and Movement Independent Gaze Interaction with Large Public Displays. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology* (Québec City, QC, Canada) *(UIST '17)*. Association for Computing Machinery, New York, NY, USA, 155–166. https://doi.org/10.1145/3126594.3126630

[52] Mohamed Khamis, Ludwig Trotter, Ville Mäkelä, Emanuel von Zezschwitz, Jens Le, Andreas Bulling, and Florian Alt. 2018. CueAuth: Comparing Touch, Mid-Air Gestures, and Gaze for Cue-Based Authentication on Situated Displays. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 4, Article 174 (Dec. 2018), 22 pages. https://doi.org/10.1145/3287052

[53] David Kim, Paul Dunphy, Pam Briggs, Jonathan Hook, John W. Nicholson, James Nicholson, and Patrick Olivier. 2010. Multi-Touch Authentication on Tabletops. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 1093–1102. https://doi.org/10.1145/1753326.1753489

[54] Yaesol Kim, Hyun Jung Kim, and Young J. Kim. 2018. Encountered-type haptic display for large VR environment using per-plane reachability maps. *Computer Animation and Virtual Worlds* 29, 3-4 (2018), 11. https://doi.org/10.1002/cav.1814

[55] Jesper Kjeldskov and Mikael B. Skov. 2014. Was It Worth the Hassle? Ten Years of Mobile HCI Research Discussions on Lab and Field Evaluations. In *Proceedings of the 16th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Toronto, ON, Canada) *(MobileHCI '14)*. Association for Computing Machinery, New York, NY, USA, 43–52. https://doi.org/10.1145/2628363.2628398

[56] Jesper Kjeldskov, Mikael B. Skov, Benedikte S. Als, and Rune T. Høegh. 2004. Is It Worth the Hassle? Exploring the Added Value of Evaluating the Usability of Context-Aware Mobile Systems in the Field. In *Mobile Human-Computer Interaction - MobileHCI 2004*, Stephen Brewster and Mark Dunlop (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 61–73.

[57] Pascal Knierim, Valentin Schwind, Anna Maria Feit, Florian Nieuwenhuizen, and Niels Henze. 2018. Physical Keyboards in Virtual Reality: Analysis of Typing Performance and Effects of Avatar Hands. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3173574.3173919

[58] Lisa Koeman. 2020. *HCI/UX Research: What methods do we use?* https://lisakoeman.nl/blog/hci-ux-research-what-methods-do-we-use/ accessed 15 September 2020.

[59] Katharina Krombholz, Heidelinde Hobel, Markus Huber, and Edgar Weippl. 2015. Advanced social engineering attacks. *Journal of Information Security and Applications* 22 (2015), 113 – 122. https://doi.org/10.1016/j.jisa.2014.09.005 Special Issue on Security of Information and Networks.

[60] Katharina Krombholz, Thomas Hupperich, and Thorsten Holz. 2016. Use the Force: Evaluating Force-Sensitive Authentication for Mobile Devices. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*. USENIX Association, Denver, CO, 207–219. https://www.usenix.org/conference/soups2016/technical-sessions/presentation/krombholz

[61] Manu Kumar, Tal Garfinkel, Dan Boneh, and Terry Winograd. 2007. Reducing Shoulder-Surfing by Using Gaze-Based Password Entry. In *Proceedings of the 3rd Symposium on Usable Privacy and Security* (Pittsburgh, Pennsylvania, USA) *(SOUPS '07)*. Association for Computing Machinery, New York, NY, USA, 13–19. https://doi.org/10.1145/1280680.1280683

[62] Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, Vol. 10. 707–710.

[63] Pedro Lopes, Sijing You, Lung-Pan Cheng, Sebastian Marwecki, and Patrick Baudisch. 2017. Providing Haptics to Walls and Heavy Objects in Virtual Reality by Means of Electrical Muscle Stimulation. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 1471–1482. https://doi.org/10.1145/3025453.3025600

[64] Ville Mäkelä, Sheikh Radiah Rahim Rivu, Saleh Alsherif, Mohamed Khamis, Chong Xiao, Lisa Marianne Borchert, Albrecht Schmidt, and Florian Alt. 2020. Virtual Field Studies: Conducting Studies on Public Displays in Virtual Reality. In *Proceedings of the 38th Annual ACM Conference on Human Factors in Computing Systems* (Honolulu, Hawaii, USA) *(CHI '20)*. ACM, New York, NY, USA, 10. https://doi.org/10.1145/3313831.3376796

[65] Karola Marky, Verena Zimmermann, Alina Stöver, Philipp Hoffmann, Kai Kunze, and Max Mühlhäuser. 2020. All in One! User Perceptions on Centralized IoT Privacy Settings. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3334480.3383016

[66] Diogo Marques, Tiago Guerreiro, Luis Carriço, Ivan Beschastnikh, and Konstantin Beznosov. 2019. Vulnerability & Blame: Making Sense of Unauthorized Access to Smartphones. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300819

[67] Florian Mathis, John Williamson, Kami Vaniea, and Mohamed Khamis. 2020. RubikAuth: Fast and Secure Authentication in Virtual Reality. In *Proceedings of the 38th Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems* (Honolulu, Hawaii, USA) *(CHI EA '20)*. ACM, New York, NY, USA, 9. https://doi.org/10.1145/3334480.3382827

[68] Florian Mathis, John Williamson, Kami Vaniea, and Mohamed Khamis. 2021. Fast and Secure Authentication in Virtual Reality using Coordinated 3D Manipulation and Pointing. *ACM Transactions on Computer-Human Interaction (ToCHI)* 28, 1 (Jan. 2021), 18. https://doi.org/10.1145/3428121

[69] A. M. Matwyshyn, A. Cui, A. D. Keromytis, and S. J. Stolfo. 2010. Ethics in security vulnerability research. *IEEE Security Privacy* 8, 2 (2010), 67–72. https://doi.org/10.1109/MSP.2010.67

[70] Michelle L. Mazurek, Saranga Komanduri, Timothy Vidas, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, Patrick Gage Kelley, Richard Shay, and Blase Ur. 2013. Measuring Password Guessability for an Entire University. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security* (Berlin, Germany) *(CCS '13)*. Association for Computing Machinery, New York, NY, USA, 173–186. https://doi.org/10.1145/2508859.2516726

[71] Lukas Mecke, Ken Pfeuffer, Sarah Prange, and Florian Alt. 2018. Open Sesame! User Perception of Physical, Biometric, and Behavioural Authentication Concepts to Open Doors. In *Proceedings of the 17th International Conference on Mobile and Ubiquitous Multimedia* (Cairo, Egypt) *(MUM 2018)*. Association for Computing Machinery, New York, NY, USA, 153–159. https://doi.org/10.1145/3282894.3282923

[72] Leap Motion. 2019. *Unity Assets for Leap Motion Orion Beta.* https://developer.leapmotion.com/unity accessed 26 August 2020.

[73] Mehdi Moussaïd, Mubbasir Kapadia, Tyler Thrash, Robert W Sumner, Markus Gross, Dirk Helbing, and Christoph Hölscher. 2016. Crowd behaviour during high-stress evacuations in an immersive virtual environment. *Journal of The Royal Society Interface* 13, 122 (2016), 20160414. https://doi.org/10.1098/rsif.2016.0414

[74] Jörg Müller, Robert Walter, Gilles Bailly, Michael Nischt, and Florian Alt. 2012. Looking Glass: A Field Study on Noticing Interactivity of a Shop Window. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New

York, NY, USA, 297–306. https://doi.org/10.1145/2207676.2207718
[75] Evangelos Niforatos, Adam Palma, Roman Gluszny, Athanasios Vourvopoulos, and Fotis Liarokapis. 2020. Would You Do It?: Enacting Moral Dilemmas in Virtual Reality for Understanding Ethical Decision-Making. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376788
[76] Borna Noureddin, Peter D Lawrence, and CF Man. 2005. A non-contact device for tracking gaze in a human computer interface. *Computer Vision and Image Understanding* 98, 1 (2005), 52–82. https://doi.org/10.1016/j.cviu.2004.07.005
[77] World Health Organization. 2020. *Coronavirus disease (COVID-19) pandemic.* https://www.who.int/emergencies/diseases/novel-coronavirus-2019 accessed 15 September 2020.
[78] Eyal Peer, Joachim Vosgerau, and Alessandro Acquisti. 2014. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior research methods* 46, 4 (2014), 1023–1031. https://doi.org/10.3758/s13428-013-0434-y
[79] Ken Pfeuffer, Melodie Vidal, Jayson Turner, Andreas Bulling, and Hans Gellersen. 2013. Pursuit Calibration: Making Gaze Calibration Less Tedious and More Flexible. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology* (St. Andrews, Scotland, United Kingdom) *(UIST '13)*. Association for Computing Machinery, New York, NY, USA, 261–270. https://doi.org/10.1145/2501988.2501998
[80] Prolific. 2020. *Quickly find research participants you can trust.* https://www.prolific.co/ accessed 15 September 2020.
[81] Susanne Putze, Dmitry Alexandrovsky, Felix Putze, Sebastian Höffner, Jan David Smeddinck, and Rainer Malaka. 2020. Breaking The Experience: Effects of Questionnaires in VR User Studies. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3313831.3376144
[82] Qualtrics. 2005. *Qualtrics Experience Managements.* https://www.qualtrics.com accessed 15 September 2020.
[83] Razer. 2020. *Razer Blade 15: NVIDIA GeForce RTX 2080.* https://www.razer.com/gb-en/gaming-laptops/razer-blade accessed 15 September 2020.
[84] Elissa M Redmiles, Yasemin Acar, Sascha Fahl, and Michelle L Mazurek. 2017. *A summary of survey methodology best practices for security and privacy researchers.* Technical Report. https://doi.org/10.13016/M22K2W
[85] Elissa M. Redmiles, Ziyun Zhu, Sean Kross, Dhruv Kuchhal, Tudor Dumitras, and Michelle L. Mazurek. 2018. Asking for a Friend: Evaluating Response Biases in Security User Studies. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security* (Toronto, Canada) *(CCS '18)*. Association for Computing Machinery, New York, NY, USA, 1238–1255. https://doi.org/10.1145/3243734.3243740
[86] Robert Rosenthal and Ralph L Rosnow. 2009. *Artifacts in behavioral research.* Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195385540.001.0001
[87] Alireza Sahami Shirazi, Peyman Moghadam, Hamed Ketabdar, and Albrecht Schmidt. 2012. Assessing the Vulnerability of Magnetic Gestural Authentication to Video-Based Shoulder Surfing Attacks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 2045–2048. https://doi.org/10.1145/2207676.2208352
[88] Johnny Saldaña. 2015. *The coding manual for qualitative researchers.* Sage.
[89] Gian-Luca Savino, Niklas Emanuel, Steven Kowalzik, Felix Kroll, Marvin C. Lange, Matthis Laudan, Rieke Leder, Zhanhua Liang, Dayana Markhabayeva, Martin Schmeißer, Nicolai Schütz, Carolin Stellmacher, Zihe Xu, Kerstin Bub, Thorsten Kluss, Jaime Maldonado, Ernst Kruijff, and Johannes Schöning. 2019. Comparing Pedestrian Navigation Methods in Virtual Reality and Real Life. In *2019 International Conference on Multimodal Interaction* (Suzhou, China) *(ICMI '19)*. Association for Computing Machinery, New York, NY, USA, 16–25. https://doi.org/10.1145/3340555.3353741
[90] Stefan Schmidt. 2009. Shall we Really do it Again? The Powerful Concept of Replication is Neglected in the Social Sciences. *Review of General Psychology* 13, 2 (2009), 90–100. https://doi.org/10.1037/a0015108
[91] Helmut Schrom-Feiertag, Volker Settgast, and Stefan Seer. 2017. Evaluation of indoor guidance systems using eye tracking in an immersive virtual environment. *Spatial Cognition & Computation* 17, 1-2 (2017), 163–183. https://doi.org/10.1080/13875868.2016.1228654
[92] Ben Shneiderman, Catherine Plaisant, Maxine Cohen, Steven Jacobs, Niklas Elmqvist, and Nicholas Diakopoulos. 2016. *Designing the User Interface: Strategies for Effective Human-Computer Interaction* (6th ed.). Pearson.
[93] Vijay Sivaraman, Hassan Habibi Gharakheili, Clinton Fernandes, Narelle Clark, and Tanya Karliychuk. 2018. Smart IoT Devices in the Home: Security and Privacy Implications. *IEEE Technology and Society Magazine* 37, 2 (2018), 71–79. https://doi.org/10.1109/MTS.2018.2826079
[94] Yunpeng Song, Yun Huang, Zhongmin Cai, and Jason I. Hong. 2020. I'm All Eyes and Ears: Exploring Effective Locators for Privacy Awareness in IoT Scenarios. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376585
[95] Anthony Steed, Francisco Ortega, Adam Williams, Ernst Kruijff, Wolfgang Stuerzlinger, Anil Batmaz, Andrea Won, Evan Rosenberg, Adalberto Simeone, and Aleshia Hayes. 2020. Evaluating Immersive Experiences During COVID-19 and Beyond. *Interactions* (2020), 62–67. https://interactions.acm.org/blog/view/evaluating-immersive-experiences-during-covid-19-and-beyond
[96] William G. Stillwell, David A. Seaver, and Ward Edwards. 1981. A comparison of weight approximation techniques in multiattribute utility decision making. *Organizational Behavior and Human Performance* 28, 1 (1981), 62 – 77. https://doi.org/10.1016/0030-5073(81)90015-5
[97] Hilary Sutcliffe. 2011. A report on responsible research and innovation. *MATTER and the European Commission* (2011), 34.
[98] Mohammad Tahaei, Kami Vaniea, and Naomi Saphra. 2020. Understanding Privacy-Related Questions on Stack Overflow. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376768
[99] Unity Technologies. 2020. *Unity Developer Documentation: Start bringing your vision to life today with the Unity real-time 3D development platform.* https://docs.unity3d.com/ScriptReference/Collider.html accessed 15 September 2020.
[100] Tobii Technology. 2020. *Tobii HTC VIVE Devkit.* https://vr.tobii.com/sdk/develop/unity/getting-started/tobii-htc-dev-kit/ accessed 26 August 2020.
[101] Lauren Thevin, Carine Briant, and Anke M. Brock. 2020. X-Road: Virtual Reality Glasses for Orientation and Mobility Training of People with Visual Impairments. *ACM Trans. Access. Comput.* 13, 2, Article 7 (April 2020), 47 pages. https://doi.org/10.1145/3377879
[102] Tobii. 2020. *Tobii Pro VR Integration.* https://www.tobiipro.com/product-listing/vr-integration/ accessed 26 August 2020.
[103] Mélodie Vidal, Andreas Bulling, and Hans Gellersen. 2013. Pursuits: Spontaneous Interaction with Displays Based on Smooth Pursuit Eye Movement and Moving Targets. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Zurich, Switzerland) *(UbiComp '13)*. Association for Computing Machinery, New York, NY, USA, 439–448. https://doi.org/10.1145/2493432.2493477
[104] Alexandra Voit, Sven Mayer, Valentin Schwind, and Niels Henze. 2019. Online, VR, AR, Lab, and In-Situ: Comparison of Research Methods to Evaluate Smart Artifacts. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, Article 507, 12 pages. https://doi.org/10.1145/3290605.3300737
[105] Emanuel von Zezschwitz, Alexander De Luca, Bruno Brunkow, and Heinrich Hussmann. 2015. SwiPIN: Fast and Secure PIN-Entry on Smartphones. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 1403–1406. https://doi.org/10.1145/2702123.2702212
[106] Emanuel von Zezschwitz, Paul Dunphy, and Alexander De Luca. 2013. Patterns in the Wild: A Field Study of the Usability of Pattern and Pin-Based Authentication on Mobile Devices. In *Proceedings of the 15th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Munich, Germany) *(MobileHCI '13)*. Association for Computing Machinery, New York, NY, USA, 261–270. https://doi.org/10.1145/2493190.2493231
[107] Roman Weiss and Alexander De Luca. 2008. PassShapes: Utilizing Stroke Based Authentication to Increase Password Memorability. In *Proceedings of the 5th Nordic Conference on Human-Computer Interaction: Building Bridges* (Lund, Sweden) *(NordiCHI '08)*. Association for Computing Machinery, New York, NY, USA, 383–392. https://doi.org/10.1145/1463160.1463202
[108] Oliver Wiese and Volker Roth. 2016. See You next Time: A Model for Modern Shoulder Surfers. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Florence, Italy) *(MobileHCI '16)*. Association for Computing Machinery, New York, NY, USA, 453–464. https://doi.org/10.1145/2935334.2935388
[109] Max Wilson, Wendy Mackay, Ed Chi, Michael Bernstein, and Jeffrey Nichols. 2012. RepliCHI SIG: From a Panel to a New Submission Venue for Replication. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI EA '12)*. Association for Computing Machinery, New York, NY, USA, 1185–1188. https://doi.org/10.1145/2212776.2212419
[110] Max L. Wilson, Ed H. Chi, Stuart Reeves, and David Coyle. 2014. RepliCHI: The Workshop II. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI EA '14)*. Association for Computing Machinery, New York, NY, USA, 33–36. https://doi.org/10.1145/2559206.2559233
[111] Christian Winkler, Jan Gugenheimer, Alexander De Luca, Gabriel Haas, Philipp Speidel, David Dobbelstein, and Enrico Rukzio. 2015. Glass Unlock: Enhancing Security of Smartphone Unlocking through Leveraging a Private Near-Eye Display. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 1407–1410. https://doi.org/10.

1145/2702123.2702316
[112] Jacob O. Wobbrock and Julie A. Kientz. 2016. Research Contributions in Human-Computer Interaction. *Interactions* 23, 3 (April 2016), 38–44. https://doi.org/10.1145/2907069
[113] Serena Zheng, Noah Apthorpe, Marshini Chetty, and Nick Feamster. 2018. User Perceptions of Smart Home IoT Privacy. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 200 (Nov. 2018), 20 pages. https://doi.org/10.1145/3274469

A APPENDIX

A.1 Semi-structured Interview Questions

Our semi-structured interviews in the usability study were loosely guided by the following questions that were asked for all three input methods: touch, mid-air, and eye gaze.

(1) General Questions
- Please tell us how you would feel using this method in public.
- Please tell us (a) what you liked; and (b) what you did not like when using this method.
- Is there anything in particular that you would like to improve in this method?
- Have you used this method previously? If yes, where?
- How did you feel when interacting with the input method? Would you define it as a positive or negative experience?

(2) VR-specific Questions [Please consider the situation where you interact with the authentication scheme you have just experienced in the real world.]
- Can you please walk us through the input method and tell us what differences may appear when using this method in the real world rather than in VR as just experienced?
- Do you think the virtual environment affected you in the way you provided input with the method?

At the end, we asked participants if they have any additional comments, questions, or suggestions.